

Improving power of genetic association studies by extreme phenotype sampling: a review and some new results

THEA BJØRNLAND^{1*}, ANJA BYE², EINAR RYENG³,
ULRIK WISLØFF², METTE LANGAAS¹

¹ *Department of Mathematic Sciences, Norwegian University of Science and Technology, Trondheim, Norway*

² *Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway*

³ *Department of Cancer Research and Molecular Medicine, Norwegian University of Science and Technology, Trondheim, Norway*

Abstract

Extreme phenotype sampling is a selective genotyping design for genetic association studies where only individuals with extreme values of a continuous trait are genotyped for a set of genetic variants. Under financial or other limitations, this design is assumed to improve the power to detect associations between genetic variants and the trait, compared to randomly selecting the same number of individuals for genotyping. Here we present extensions of likelihood models that can be used for inference when the data are sampled according to the extreme phenotype sampling design. Computational methods for parameter estimation and hypothesis testing are provided. We consider methods for common variant genetic effects and gene-environment interaction effects in linear regression models with a normally distributed trait. We use simulated and real data to show that extreme phenotype sampling can be powerful compared to random sampling, but that this does not hold for all extreme sampling methods and situations.

Key words: GWAS, gene-environment interactions, extreme phenotype sampling, outcome-dependent sampling, selective genotyping, the HUNT study

1 Introduction

Extreme phenotype sampling (EPS) is an outcome-dependent sampling design for genetic association studies. For this design, individuals with high or low values of a particular continuously measurable phenotype (trait) are genotyped. When the number of individuals that can be genotyped is limited, such samples are assumed to give good power to detect associations between genetic variants and the trait. Random sampling is the most relevant competing design. Extreme samples must be analyzed with statistical methods that properly account for the sampling bias, and these methods are not trivial. Random samples can be analyzed with readily available standard methods and are therefore preferable when it comes to data analysis. However, low statical power is an important issue in genetic association studies (Sham and Purcell [2014], Hirschhorn et al. [2002]). Here we attempt to answer whether, when and to what degree extreme sampling is more powerful than random sampling in genetic association studies. For this purpose we have extended relevant likelihood methods for parameter estimation and hypothesis testing.

We consider genetic association studies where the aim is to detect common genetic variants that are associated with some trait, and also to quantify associations. We consider biallelic single-nucleotide polymorphism (SNP) data. For a particular SNP, the observed genotype for an individual is either aa , aA or AA , where A represents the minor-allele in the population. We consider additive genetic models so that the genotype is coded as 0, 1 or 2 according to the number of copies of the minor-allele. In a *genome-wide association study* (GWAS) the observed genotypes of selected

SNPs along the genome, so-called genetic markers, are tested for association with a phenotype in a sample of individuals. The number of SNPs is large ($\sim 10^6$). The purpose of such studies is to detect regions in the genome that are associated with the phenotype. In what we will refer to as a *candidate SNP study*, a small collection SNPs are analyzed. These SNPs can for example be chosen based on results from studies in other populations, or from studies of related traits and diseases. Then the focus is on replication and effect size estimation rather than detection.

Due to high genotyping costs and low statistical power to detect significant associations between genetic variants and complex traits, selective genotyping has been proposed as a strategy for achieving good statistical power under sample size limitations. Genotyping only the phenotypically extreme individuals was proposed for mapping quantitative trait loci (QTL) in experimental organisms (Lebowitz et al. [1987], Lander and Botstein [1989]). The methodology was further developed for linkage disequilibrium mapping of QTLs by Darvasi and Soller [1992], Slatkin [1999], Chen et al. [2005], Wallace et al. [2006], among others. Power studies have been performed by Van Gestel et al. [2000] and Xing and Xing [2009]. Most methods were not fully efficient because they involved discretizing the continuous trait, discarding individuals who were homozygous for the minor-allele or not accounting for the biased sampling. To that effect Huang and Lin [2007] proposed likelihood methods that made full use of the data and accounted for the selective genotyping design. Recently, likelihood methods for multivariate trait-dependent sampling has also been considered (Lin et al. [2013], Tao et al. [2015]). Selective genotyping has also recently been proposed in studies of rare genetic variants (Li et al. [2011], Guey et al. [2011], Barnett et al. [2013]). The extreme phenotype sampling design can be considered a special case of the outcome-dependent sampling design described by Zhou et al. [2002] and Weaver and Zhou [2005].

The power, limitations and practical utility of extreme phenotype sampling as compared to random sampling in modern GWAS and candidate SNP studies has in our opinion not been sufficiently characterized, and this might explain why the design has not been much used. Huang and Lin [2007] considered regression models for continuous phenotypes and developed likelihood models for extreme sampling data for the special case where only one covariate (the genetic variant) was included in the regression model. The power of two different extreme sampling designs was estimated, but not compared to random sampling. Using the asymptotic distribution of the score test statistic under the alternative hypothesis Tang [2010] showed that the methods by Huang and Lin [2007] theoretically can give better power than a random sampling design in the special case with no non-genetic covariates. Zhou et al. [2002] showed that their likelihood method for the outcome-dependent sampling design yielded more efficient parameter estimates than would be obtained using a simple random sample of the same size. We extend the likelihood methods for the EPS-design to include non-genetic (environmental) explanatory variables as well as gene-environment interaction terms. As experienced by us in a study of gene-environment interactions and obesity using an extreme sampling design this additional model complexity is necessary for application purposes [Bjørnland et al., 2016]. We assess the statistical power and other properties of our methods using both simulated and real data. The data set is from the HUNT study (Helseundersøkelsen i Nord-Trøndelag) which comprises health information on the population of Nord-Trøndelag county, Norway [Krokstad et al., 2013]. We use data from a GWAS on the trait *maximum oxygen uptake* based on the HUNT Fitness study [Aspenes et al., 2011]. All our computational methods for parameter estimation and hypothesis testing under the EPS-design are available in our R-package.

2 Models and methods

We consider complex traits that are continuously measurable and can be assumed to be normally distributed in the population. Let the i th individual in a population (or large random sample) of size N have observed trait value y_i , environmental variables $\mathbf{x}_{ei}^T = (x_{ei1}, \dots, x_{eid})$, and SNP genotypes $\mathbf{x}_{gi}^T = (x_{gi1}, \dots, x_{gim})$. Assume that the continuous trait Y_i can be modeled by the linear regression model

$$Y_i = \alpha + \mathbf{x}_{ei}^T \boldsymbol{\beta}_e + \mathbf{x}_{gi}^T \boldsymbol{\beta}_g + (\mathbf{x}_{ei} \mathbf{x}_{gi})^T \boldsymbol{\beta}_{eg} + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d } \mathcal{N}(0, \sigma^2), i = 1, \dots, N, \quad (1)$$

where $\mathbf{x}_{ei} \mathbf{x}_{gi}$ represents a vector of interactions between some environmental and genetic covariates, e.g. $(\mathbf{x}_{ei} \mathbf{x}_{gi})^T = (x_{eij} x_{gik}, x_{eij} x_{gil})$ for some $j \in \{1, \dots, d\}$ and $k, l \in \{1, \dots, m\}$, $k \neq l$.

We consider GWAS and candidate-SNP studies. In genome-wide studies we analyze a large number of SNPs, and each SNP is tested separately for association with the phenotype ($H_0 : \beta_{gk} = 0$ against $H_1 : \beta_{gk} \neq 0$, $k = 1, \dots, m$) without any interaction effects. P -values are compared to a significance threshold that is determined by the multiple testing burden. In candidate-SNP studies, a few selected SNPs are studied, and the aim is to test for associations between the phenotype and all or some of the genetic variables ($H_0 : \beta_g = 0$ against $H_1 : \beta_g \neq 0$, or $H_0 : \beta_{gk} = 0$ against $H_1 : \beta_{gk} \neq 0$, $k = 1, \dots, m$), and to obtain parameter estimates and confidence intervals for the genetic effects. Furthermore, it is also of interest to study gene-environment interaction effects ($H_0 : \beta_{eg} = 0$ vs $H_1 : \beta_{eg} \neq 0$).

Assume that due to financial or other limitations, only $n < N$ individuals can be genotyped. We define extreme phenotypes as observations by $y_i < c_l$ and $y_i > c_u$, for some cut-off values c_l and c_u such that $c_l < c_u$. In the sample of size N , the cut-offs can be chosen such that we sample the n most extreme individuals (e.g. $n/2$ from each tail), or we could choose less extreme cut-offs and thereafter draw n individuals from the extremes. The methods presented here apply to both situations. Our main analysis concerns two different extreme sampling designs which we refer to as the EPS-only and the EPS-full sampling designs. These are extensions of the designs discussed by Huang and Lin [2007]. In the EPS-only design, observations of any variable (y_i , \mathbf{x}_{ei} or \mathbf{x}_{gi}) are available *only* for extreme phenotype individuals. In the EPS-full design, observations of the phenotype (y_i) and environmental covariates (\mathbf{x}_{ei}) are available for the *full* population, while genetic variants (\mathbf{x}_{gi}) are only observed for the extremes. Let \mathcal{C} denote the set of indexes of the n extreme phenotype individuals.

2.1 EPS-only

For the EPS-only sampling design the observations ($y_i, \mathbf{x}_{ei}, \mathbf{x}_{gi}$) are available for all individuals $i \in \mathcal{C}$, i.e. all extreme-phenotype individuals. We consider two different statistical methods for this design; the EPS-only binary method and the EPS-only (continuous) method.

The first method treats the lower and upper extremes as binary responses and we refer to this method as the EPS-only binary method. The second method takes the continuity of the trait into account, and we refer to this as the EPS-only (continuous) method.

2.1.1 EPS-only binary

Were we treat the lower and upper phenotypic extremes as a binary response. One method for analysis of EPS-only data is to test whether allele frequencies are significantly different between the two extreme tails, for example by using contingency tables. We use a logistic regression model in order to include environmental covariates and gene-environment interaction terms. Define the variable Y_{di} such that $Y_{di} = 0$ if $Y_i < c_l$, $Y_{di} = 1$ if $Y_i > c_u$. Let π_i denote the probability $P(Y_{di} = 1; \mathbf{x}_{ei}, \mathbf{x}_{gi} | Y_i < c_l \cup Y_i > c_u)$, such that $1 - \pi_i = P(Y_{di} = 0; \mathbf{x}_{ei}, \mathbf{x}_{gi} | Y_i < c_l \cup Y_i > c_u)$. A logistic regression model

$$\text{logit}(\pi_i) = a + \mathbf{x}_{ei}^T \mathbf{b}_e + \mathbf{x}_{gi}^T \mathbf{b}_g + (\mathbf{x}_{ei} \mathbf{x}_{gi})^T \mathbf{b}_{eg}, \quad (2)$$

can be fitted to the dichotomized extreme sample data. Under the two-sided hypothesis $H_0 : \beta_g = 0$ in the linear regression model (1), there is no difference between allele frequencies in the lower and upper extremes, which in the dichotomized sample can be tested by the two-sided hypothesis $H_0 : \mathbf{b}_g = 0$, and similarly for gene-environment interactions. Hypothesis tests for this model can be done using standard methods for logistic regression. Note that the parameters of the logistic regression model (2) are directly dependent upon the choice of c_l and c_u and comparison of results between studies must be done cautiously.

2.1.2 EPS-only (continuous)

A likelihood model for EPS-only samples with a continuous response was proposed by Huang and Lin [2007], then called the conditional likelihood. Here, we extend this likelihood to include environmental covariates and gene-environment interactions and develop the score test. Let Y_{ci} denote a random variable from the extremes of the distribution of Y_i . Then $F_{Y_{ci}}(y) = P(Y_i \leq$

$y|Y_i < c_l \cup Y_i > c_u$) and the probability density can be derived accordingly (see Appendix A). The likelihood for the EPS-only sample is

$$L = \prod_{i \in \mathcal{C}} \frac{\frac{1}{\sigma} \phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right)}{1 - \Phi\left(\frac{c_u - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right) + \Phi\left(\frac{c_l - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right)}, \quad (3)$$

where $\Phi()$ is the cumulative probability distribution and $\phi()$ is the density function of the standard normal distribution, and $\mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg}) = \alpha + \mathbf{x}_{ei}^T \beta_e + \mathbf{x}_{gi}^T \beta_g + (\mathbf{x}_{ei} \mathbf{x}_{gi})^T \beta_{eg}$.

We have implemented a quasi-Newton numerical optimization method to obtain likelihood estimates. We obtain approximate $(1 - \alpha)100\%$ confidence intervals for some parameter β_j by $\left[\hat{\beta}_j - z_{\alpha/2} \sqrt{(I_o^{-1})_{\beta_j, \beta_j}}, \hat{\beta}_j + z_{\alpha/2} \sqrt{(I_o^{-1})_{\beta_j, \beta_j}}\right]$, where $\hat{\beta}_j$ is the maximum likelihood estimate and $z_{\alpha/2}$ is such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. The observed information matrix I_o is estimated in the optimization.

We are interested in the two-sided hypothesis tests $H_0 : \beta_g = 0$, $H_0 : \beta_{gk} = 0$ for $k = 1, \dots, m$, and $H_0 : \beta_{eg} = 0$. In the GWAS setting we strongly prefer a computationally efficient test due to the large number of tests. Using the score test, the null model can be fitted once and for GWAS in particular, the test is then computationally fast compared to the likelihood ratio test which requires model fitting under all m alternative hypotheses. For the EPS-only likelihood (3), we have derived a closed form expression for the score test statistic (see Appendix A.1). The expression is mathematically complex and requires some tedious algebra, but the reward is significant computational efficiency. Tang [2010] showed that in the most simple model ($y = \alpha + x_g \beta_g + \varepsilon$) the score test statistic for $H_0 : \beta_g = 0$ derived from the continuous EPS-only likelihood is equivalent to the the score test statistic derived from the likelihood for a normal linear regression model. In Appendix A.1, we show that this also holds for testing the two-sided hypothesis $H_0 : \beta_g = 0$ in the model $y = \alpha + \mathbf{x}_g^T \beta_g + \varepsilon$, but not when other covariates (\mathbf{x}_e) are included in the null model.

2.2 EPS-full

The EPS-full sample consists of observations $(y_i, \mathbf{x}_{ei}, \mathbf{x}_{gi})$ for all $i \in \mathcal{C}$, and observations (y_i, \mathbf{x}_{ei}) for all $i \notin \mathcal{C}$. In other words, \mathbf{x}_{gi} is missing for all individuals $i \notin \mathcal{C}$. The missing observations are missing at random (MAR) because the observations are not missing due to the unobserved \mathbf{x}_{gi} , but rather due to the observed value of the phenotype y_i . We consider both a likelihood based method and multiple imputation for this sample.

2.2.1 EPS-full likelihood

Under MAR the likelihood that ignores the missing-mechanism is valid for likelihood inference from the frequentist perspective [Little and Rubin, 2002, page 120]. It is necessary to specify or estimate the distribution of the variables that are missing. We assume in the most general case that \mathbf{X}_g is dependent upon some (or all) of the covariates in \mathbf{X}_e , denoted by \mathbf{X} . We assume that the sample space of \mathbf{X} is discrete with elements \mathbf{x}_j , $j = 1, \dots, J$. We let \mathbf{x}_{gk} , $k = 1, \dots, K$ denote elements of the sample space of \mathbf{X}_g and we assume that $\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j$ can take any value in this sample space for all j , albeit with different probabilities. The likelihood for the EPS-full sample is an extension of the so-called full likelihood by Huang and Lin [2007] and is derived in Appendix B. The EPS-full likelihood is

$$L = \prod_{i \in \mathcal{C}} \frac{1}{\sigma} \phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right) \sum_{j=1}^J f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gi}) I(\mathbf{x}_i = \mathbf{x}_j) \cdot \prod_{i \notin \mathcal{C}} \sum_{k=1}^K \frac{1}{\sigma} \phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gk}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right) \sum_{j=1}^J f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) I(\mathbf{x}_i = \mathbf{x}_j), \quad (4)$$

where $\phi()$ is the density function of the standard normal distribution and $f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}}(\mathbf{x}_g)$ is the probability mass function of the SNPs. Note that \mathbf{x}_{ei} to all intents and purposes is a constant

vector in the EPS-full model, while Y_i and \mathbf{X}_{gi} are random. The missing-structure is determined by the choice of c_u and c_l . These parameters may be chosen such that the distinctness condition for the ignorability principle [Little and Rubin, 2002, page 119] does not hold (e.g. choosing as cut-offs the quantiles of the empirical distribution of Y). Then the likelihood that ignores the missing mechanism (4) is still valid, but not fully efficient (e.g. there is some information about the distribution of Y in the empirical quantiles). We prefer to ignore the missing-mechanism so that the likelihood can be used for various trait-dependent sampling designs.

In some models, the distribution of the genotypes of the SNPs does not depend on the value of the non-genetic covariates and $f_{\mathbf{X}_g|\mathbf{X}=\mathbf{x}_j}(\mathbf{x}_g) = f_{\mathbf{X}_g}(\mathbf{x}_g)$. However, in the presence of confounding effects (e.g. population stratification), the distribution of the genotypes will differ in subsets of the sample, and the assumed genotype distribution must account for this. The sample space for each SNP is $\{0, 1, 2\}$, and we assume that $P(X_{gk} = 0|\mathbf{X} = \mathbf{x}_j) = p_{0kj}$, $P(X_{gk} = 1|\mathbf{X} = \mathbf{x}_j) = p_{1kj}$ and $P(X_{gk} = 2|\mathbf{X} = \mathbf{x}_j) = 1 - p_{0kj} - p_{1kj}$. A more general distribution of the missing covariate in a similar likelihood model has been considered in the literature (Lawless et al. [1999], Ibrahim et al. [2005]) but we have used this simple SNP property for computational purposes. If Hardy-Weinberg equilibrium is assumed [Ziegler et al., 2010, page 39], then we have $P(X_{gk} = 0|\mathbf{X} = \mathbf{x}_j) = (1 - q_{kj})^2$, $P(X_{gk} = 1|\mathbf{X} = \mathbf{x}_j) = 2q_{kj}(1 - q_{kj})$ and $P(X_{gk} = 2|\mathbf{X} = \mathbf{x}_j) = q_{kj}^2$, where q_{kj} is the minor allele frequency. In GWA studies where SNPs are tested one at a time a joint distribution of the genetic variants is not relevant. In candidate-SNP studies, the SNPs that are considered are typically such that the genotype distributions can be assumed to be statistically independent (e.g. SNPs from different genes or chromosomes), and one can define the joint distribution as the product of the marginal distributions. Otherwise, a multivariate multinomial distribution can be used.

As for the EPS-only likelihood, we have implemented a quasi-Newton numerical optimization method to obtain maximum likelihood estimates and confidence intervals. Kenward and Molenberghs [1998] showed that the observed information matrix can be used as an estimate of the true information matrix for likelihoods that ignore the missing-mechanism.

The EPS-full likelihood (4) is valid for direct-likelihood inference [Rubin, 1976], which ensures validity of the likelihood ratio test. For the score test, the asymptotic variance of the score vector depends on the missing-mechanism. We therefore derived the score test by using the observed information matrix under the null to approximate the asymptotic variance of the score vector, which is a valid estimate of the true information matrix [Kenward and Molenberghs, 1998]. See Appendix B.1 for the derivation of the EPS-full score test for $H_0 : \beta_g = 0$. Again, the derivation is complex, but the test is computationally efficient. We used the result of Derkach et al. [2015] to obtain the simplest closed form expression for the score test statistic. If covariates with a missing-structure are present in the null model, for example when testing $H_0 : \beta_{eg} = 0$, a similar (relatively) simple closed form expression for the score test statistic cannot be attained, and we have implemented the likelihood ratio test for this purpose.

2.2.2 EPS-full with multiple imputation

Multiple imputation (MI) is a tool that can be used for parameter estimation and hypothesis testing for the linear model (1) when covariates are MAR. We use the method of multivariate imputation by chained equations (MICE), also known as fully conditional specification [van Buuren, 2007]. Broadly speaking, the method imputes the missing genotypes by sampling from an empirical conditional distribution of the genetic variant for individual i , given all other observations in the sample. This is repeated to create m_{MI} different data sets in which model inference is performed. Lastly, parameter estimates or test statistics are pooled into one estimate. The MICE method is readily available in many statistical softwares, and we have used the R-package `mice` [Buuren and Groothuis-Oudshoorn, 2011]. Our aim is not to develop a specific multiple imputation method for extreme sampling in genetic association studies, but rather to compare our proposed EPS-full likelihood method to an existing inference method for missing data problems. We refer to this method as the EPS-full MI method.

3 Simulation study

Using simulated data, we compare the performance of the EPS methods (EPS-only binary, EPS-only, EPS-full, EPS-full MI), with results from the full sample and a random sample. We generated a full data set of size N , and selected $n < N$ individuals for genotyping under extreme and random sampling. We set c_l and c_u such that the set \mathcal{C} consisted of the $n/2$ lowest and $n/2$ highest extremes of the empirical phenotype distribution. For the EPS-only design we then discarded all information on non-extremes ($i \notin \mathcal{C}$). For the EPS-full design, we discarded the genotype information for all individuals $i \notin \mathcal{C}$. We considered the following simulations models;

$$Y = \alpha + \beta_{e1}x_{e1} + \beta_{e2}x_{e2} + \beta_g x_g + \varepsilon, \quad (5)$$

$$Y = \alpha + \beta_{e1}x_{e1} + \beta_{e2}x_{e2} + \beta_g x_g + \beta_{e1g}x_{e1}x_g + \varepsilon, \quad (6)$$

$$Y = \alpha + \beta_{e1}x_{e1} + \beta_{e2}x_{e2} + \beta_g x_g + \beta_{e2g}x_{e2}x_g + \varepsilon. \quad (7)$$

The non-genetic covariate x_{e1} is a Bernoulli(0.4) random variable, x_{e2} is a $N(2, 1)$ random variable and the genetic marker x_g is a multinomially distributed random variable taking values (0, 1, 2) with probabilities (0.49, 0.42, 0.09). These probabilities were generated by assuming Hardy-Weinberg equilibrium and a minor allele frequency $q = 0.3$. For multiple imputation we set $m_{MI} = 10$. The parameter values that we used are given in Table 1. With these parameter choices, the environmental covariates x_{e1} and x_{e2} were much more important than the genetic covariate for describing the response (R^2 from fitting the regression model with and without the genetic covariates varied minimally). This choice was motivated by the assumption that environmental variables are more important for predicting a complex trait, compared to the genotype of a common genetic variant (Darvasi and Soller [1992], Manolio et al. [2009]).

Parameter	Value
N	5000
n	$N/2$
α	50
β_{e1}	10
β_{e2}	5
β_g	0.5
β_{e1g}	1
β_{e2g}	0.5
σ	6
q	0.3

Table 1: Specification of parameters used in simulated data sets for simulation models (5), (6) and (7).

3.1 Main effects model

3.1.1 Parameter estimation

We generated $R = 10\,000$ data sets using simulation model (5) and obtained the maximum likelihood estimate $\hat{\beta}_g$ for the different designs and methods. We estimated the mean squared error (MSE) by $\frac{\sum_{r=1}^R (\hat{\beta}_{gr} - \beta_g)^2}{R}$. For the full and random sample we used the function `lm()` in R to obtain parameter estimates. For the EPS-only and EPS-full likelihood methods we used functions provided in our R-package, and for the EPS-full MI method we used the R-package `mice` [Buuren and Groothuis-Oudshoorn, 2011]. Results for different sample sizes N and genotyped sample size $n = N/2$ are presented in Table 2. Compared to results from the full sample, the EPS-full likelihood method gave the lowest MSE, while the random sample, the EPS-only sample and the EPS-full MI method had similar and slightly higher MSE. This relationship was the same for increasing values of the full sample size N , and the MSE decreased as N increased for all methods.

N	Full	Random	EPS-only continuous	EPS-full likelihood	EPS-full MI
1000	0.085	0.170	0.163	0.128	0.162
3000	0.028	0.056	0.055	0.043	0.057
5000	0.017	0.034	0.034	0.027	0.034
7000	0.012	0.024	0.023	0.018	0.022

Table 2: Estimated mean squared error for the coefficient of the genetic variant (β_g) in simulation model (5) for different sample sizes N with all other parameters held fixed at the values described in Table 1.

3.1.2 Power

We estimated the power to detect a non-zero genetic effect ($\beta_g \neq 0$) in simulation model (5). We generated $R = 10\,000$ data sets under the alternative hypothesis, and tested $H_0 : \beta_g = 0$ against $H_1 : \beta_g \neq 0$ for all designs and methods. For the full, random and EPS-only binary models we used the score test provided in the R-package `statmod` [Dunn and Smyth, 1996]. Power was estimated at a 5% significance level by $\frac{\sum_{r=1}^R I_{(0,0.05)}(p_r)}{R}$, where $I_{(0,0.05)}(p_r)$ is the indicator function and p_r is the p -value in the r th simulated data set.

Power estimates for different values of β_g , σ , the minor allele frequency q and the full sample size N are presented in Table 3. Expectedly, all methods had highest power when the variance σ^2 was low, the parameter β_g was large, the minor allele frequency q was high and the sample size N was large. For all values of β_g , the EPS-only binary model had lowest power. The EPS-full likelihood method had higher estimated power than all the alternatives. The EPS-only continuous model was the second most powerful but only slightly better than random sampling. The EPS-full MI method was slightly less powerful than random sampling. When the variance was high all EPS-models performed better than random sampling. For low variance, the random sample was more powerful than the EPS-only binary and EPS-full MI methods. For all q the EPS-only binary model performed worst, the random sample, EPS-full MI and EPS-only method were similar, while the EPS-full likelihood method was most powerful. We observe similar results for different values of N . Overall, the EPS-full likelihood model performed notably better than all alternatives. This is in contrast to the results by Huang and Lin [2007] who found that in models with no environmental covariates the EPS-full and EPS-only likelihood methods performed similarly.

We also considered power as a function of n - the number of genotyped individuals. We set the parameters of the simulation model (5) such that the power to detect non-zero β_g was approximately 80% in the full sample ($N = 5000$). All parameters were as in Table 1 except that we set $\sigma = 8$. In $R = 10\,000$ simulated data sets we considered n ranging from 1000 to 5000 in increments of 500. Power simulation results are presented in Figure 1. If we for example wanted to design a study with 70% power, we see that we would have to genotype approximately 3000 individuals for the EPS-full likelihood method, 3500 for EPS-only likelihood method and 4000 for a random sample and the EPS-full MI method. The EPS-only binary model never achieves 70% power in this scenario. In EPS-only binary we test whether the genotype frequencies are significantly different between the upper and lower extreme groups. As n increases, these two groups become more similar. Therefore, the power of the EPS-only binary model does not converge towards the power of the full sample as n increases.

3.1.3 Computational efficiency

The computational time for testing $H_0 : \beta_g = 0$ in 100 simulated data sets was 1.7 seconds for full samples, 1.1 seconds for random samples, 2.1 seconds for the EPS-only sample with the binary method, 18.37 seconds for the EPS-only continuous method, 1.7 seconds for the EPS-full likelihood method and 14 minutes for the EPS-full MI method. All computations were performed using R version 3.3.1 [R Core Team, 2016] on a personal computer (MacBook Air (13", Early 2014) with 1.7 GHz Intel Core i7-4650U with 4 MB cache). The score test for the EPS-full likelihood method is computationally efficient because under the null hypothesis we fit a linear model to the full sample (no missing variables). The score test for EPS-only is slower because model fitting under

		Full	Random	EPS-only binary	EPS-only continuous	EPS-full likelihood	EPS-full MI
β_g	0.3	62.47	36.86	24.37	37.67	46.09	30.25
	0.5	96.97	76.76	57.16	78.36	87.36	71.72
	0.7	99.95	96.39	84.86	96.95	99.03	95.26
σ	4	99.99	98.25	55.17	94.72	98.99	94.48
	6	96.97	77.42	57.94	79.77	87.81	73.53
	8	81.30	52.89	47.29	60.85	68.34	51.61
	10	63.20	36.60	38.42	47.60	52.83	38.97
q	0.1	70.76	42.85	30.13	44.89	54.51	37.63
	0.2	91.80	65.61	47.27	68.42	77.85	60.92
	0.3	96.47	76.60	56.91	78.69	87.42	72.83
N	1000	39.48	22.09	17.05	24.18	28.06	20.12
	3000	83.98	55.68	39.14	57.37	67.49	51.28
	5000	96.79	77.09	58.30	79.17	87.66	73.62
	7000	99.53	88.96	72.26	90.32	96.03	86.41

Table 3: Estimated power to detect a non-null genetic effect ($H_0 : \beta_g = 0$) in simulation model (5) for different values of β_g , σ , minor allele frequency q and full sample size N . For each parameter that varied, all other parameters were held fixed at the values described in Table 1.

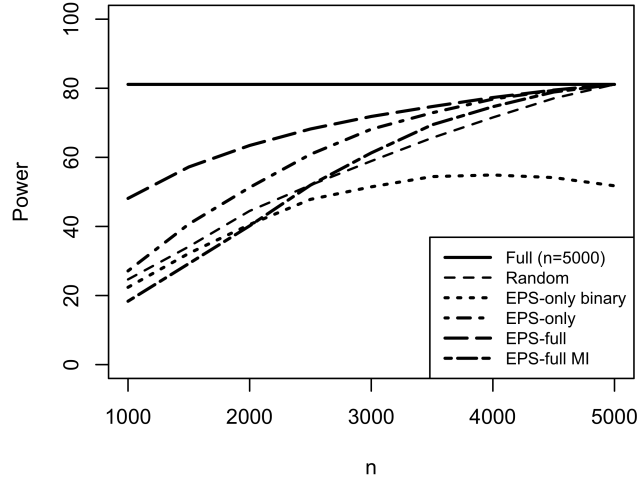


Figure 1: Estimated power to detect a non-null genetic effect ($H_0 : \beta_g = 0$) in simulation model (5) for increasing number of genotyped individuals (n), compared to the power for the full sample where $n = N = 5000$.

the null requires numerical optimization of the EPS-only likelihood. We observe that the EPS-full MI method that we have used is computationally slow.

3.2 Gene-environment interaction models

For simulation models (6) and (7) we estimated power for the two-sided tests of $H_0 : \beta_{e1g} = 0$ and $H_0 : \beta_{e2g} = 0$ at a 5% significance level, for different values of β_{e1g} and β_{e2g} . Model parameters were as in Table 1. For the EPS-full MI method with interactions, we imputed using the passive method described in Buuren and Groothuis-Oudshoorn [2011] (Section 3.4). In simulation model (6) there is an interaction between a genetic variant and a *binary* environmental variable. From the power estimates for increasing values of β_{e1g} (Table 4) we see that the EPS-only binary method and EPS-full MI method performed poorly in this setting, and furthermore that the EPS-only continuous method was similar to a random sample. The EPS-full likelihood method had the highest estimated power. In simulation model (7) there is an interaction between a genetic variant and *continuous* environmental variable. Also here, the EPS-full method performed the best in our simulations, while the EPS-only method was slightly better than a random sample. The EPS-only binary and EPS-full MI methods again had the lowest estimated power.

		Full	Random	EPS-only continuous	EPS-only binary	EPS-full likelihood	EPS-full MI
β_{e1g}	0.8	84.13	56.45	55.95	33.21	62.25	18.51
	1.0	96.21	75.45	75.07	47.32	81.21	33.00
	1.2	99.34	89.45	88.13	62.44	93.00	52.73
β_{e2g}	0.4	85.54	63.57	57.72	27.95	69.46	29.97
	0.6	99.51	92.96	89.73	50.66	96.14	73.08
	0.8	100.0	99.74	99.16	77.04	99.94	97.57

Table 4: Estimated power to detect a non-null gene-environment interaction effect in simulation model (6) ($H_0 : \beta_{e1g} = 0$) and in simulation model (7) ($H_0 : \beta_{e2g} = 0$). All parameters other than β_{e1g} and β_{e2g} were held fixed at the values described in Table 1.

4 Application to data from the HUNT study

We assessed the extreme sampling methods by application in a relevant data set. Our data set comes from the HUNT Fitness study [Aspenes et al., 2011]. A genome-wide association study for maximum volume uptake of oxygen (VO_2) has been performed by co-authors Anja Bye, Einar Ryeng and Ulrik Wisløff. The participants in this study represent a full sample (all participants were genotyped) and the trait VO_2 can be assumed normally distributed in the population. In the original study, the regression model

$$VO_2 = \alpha + \beta_{e2}x_{age} + \beta_{e1}x_{sex} + \beta_{e3}x_{PA} + \beta_{gk}x_{gk} + \varepsilon$$

was used to test $H_0 : \beta_{gk} = 0$ against $H_1 : \beta_{gk} \neq 0$ for $k = 1, \dots, m$. The non-genetic covariates were age, sex and physical activity (PA). In the original study, the main associations between genetic variants and oxygen uptake was found in chromosome 1. We therefore used this chromosome to illustrate our methods and the number of SNPs to test was then $m = 11098$. We here excluded all participants with missing non-genetic covariates (age, sex, physical activity). The full sample size was then $N = 2802$.

We considered the fictitious situation where we could only afford to genotype half of the full sample. We performed tests for association across chromosome 1 using the full data set, a random sample of size $N/2$ and extreme samples (lower and upper quartiles). The full data set also had some missing genotype observations, as is common in genetic association studies. For the full, random and EPS-only samples, we imputed the mean genotype. For the EPS-full model we assume that the sample space of each genetic variant is $\{0, 1, 2\}$ and mean imputation would be at odds with

this assumption. However, when genotypes are missing at random (MAR) or missing completely at random (MCAR), the EPS-full likelihood (4) is valid. We assumed MCAR for the initially missing genotypes, and MAR for the genotypes that were missing due to extreme sampling, and then no imputation was necessary. For the EPS-full MI method we considered one SNP at a time and imputation was then based only on VO_2 , sex, age and physical activity.

The Manhattan plot for each sampling method is shown in Figure 2. A Manhattan plot is a plot of $-\log_{10}(p)$, p being the p -value from the two-sided test of $H_0 : \beta_g = 0$, against the position of the SNP on the genome. Such plots are typical for GWAS, and regions on the genome where there is a peak in $-\log_{10}(p)$ -values are considered to be of further interest. In the Manhattan plot of the full sample we see that such a peak appears in chromosome 1. Furthermore, we see a very similar result for the EPS-full sample for both the likelihood and MI methods. The peak can also be distinguished in the EPS-only model, but not when using the EPS-only binary method, nor in the random sample. Genotyping only $n = N/2$ extreme phenotype individuals in this study could have been sufficient to detect the same region that was found when genotyping all N individuals. We note that the performance of EPS-full MI method was much better here than using simulated data. We also note that the extreme sample as analyzed by the EPS-only (continuous) method is clearly better than the "unlucky" random sample drawn here.

For the top finding in the full sample we also estimated β_g in the full, random sample and extreme samples. Parameter estimates were 2.21 (95% CI = (1.24, 3.17), $p = 7.9 \cdot 10^{-6}$) in the full sample, 1.23 (95% CI = (-0.16, 2.98), $p = 0.08$) in the random sample, 2.85 (95% CI = (1.51, 4.19), $p = 3.4 \cdot 10^{-5}$) in the EPS-only sample, 2.74 (95% CI = (1.55, 3.92), $p = 3.0 \cdot 10^{-6}$) in the EPS-full sample with the maximum likelihood method, and 3.02 (95% CI = (1.57, 4.47), $p = 4.0 \cdot 10^{-5}$) in the EPS-full sample using multiple imputation. The results from the EPS-full likelihood method was closest to the results of the full sample. All EPS-methods were slightly biased upwards.

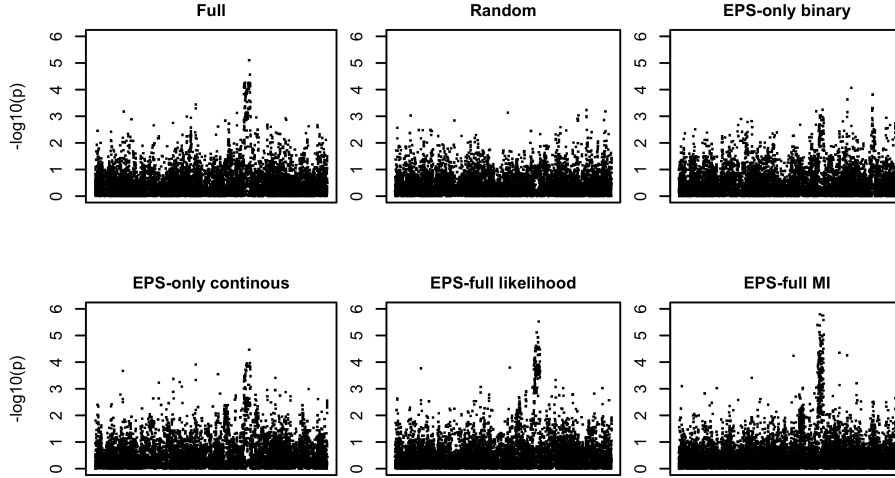


Figure 2: Manhattan-plot for testing each SNP in chromosome 1 against VO_2 for the full model (all N study participants analyzed), a random sample ($n = N/2$ randomly drawn participants analyzed), an EPS-only sample ($n = N/2$ most extreme participants analyzed) using the EPS-only binary method and the EPS-only continuous method, and an EPS-full sample (N participants analyzed of which only $n = N/2$ most extreme participants had observed genotypes) using the EPS-full likelihood and EPS-full MI method.

5 Other extreme sampling methods and designs

5.1 Gene-environment interactions by an extreme exposure sampling design

With the purpose to study gene-environment interactions, Boks et al. [2007] proposed an extreme-exposure sampling (EES) design. In this design, individuals with extreme values of the relevant environmental covariate are sampled for genotyping. We refer to the EES-only design as a sample where we only observe individuals with extreme exposure values, as considered by Boks et al. [2007]. Furthermore, we considered an EES-full design where extreme exposure individuals are genotyped but where non-genetic information is also available for all other individuals. EES-only data can be analyzed using standard linear regression methods, while EES-full data can be analyzed with the EPS-full likelihood (4). The missing-mechanism is MCAR since \mathbf{x}_{ei} is not a random variable in these models. We set up a simulation study to compare the performance of the EES designs to the EPS designs for gene-environment interaction effects. We consider the simulation model (7), which has an interaction between a continuous environmental covariate and a genetic covariate. Parameter values were as in Table 1 and we simulated $R = 10\,000$ data sets. For different values of β_{e2g} , the estimated power of the different models are presented in Table 5. We see that the EES-only and EES-full samples had almost identical estimated power in our simulations, and both performed better than the EPS-only and EPS-full samples. A discussion on why EES-only and EES-full have almost identical power under the MCAR criterion is given in Appendix D. For the specific purpose of studying an interaction between a continuously measurable environmental exposure and genetic variants, the extreme exposure design seems powerful and the data set is simple to analyze.

β_{e2g}	Full	EPS-only	EPS-full	EES-only	EES-full
0.3	51.78	32.41	35.98	48.95	49.12
0.4	76.70	52.12	57.55	73.91	74.04
0.5	91.94	71.95	78.06	89.67	89.93

Table 5: Estimated power to detect a non-null gene-environment interaction effect in simulation model (7) ($H_0 : \beta_{e2g} = 0$) for extreme phenotype sampling (EPS) and extreme exposure sampling (EES). All parameters other than β_{e2g} were held fixed at the values described in Table 1.

5.2 Combining extreme and random sampling

The outcome-dependent sampling design (Zhou et al. [2002], Weaver and Zhou [2005]) is a generalization of extreme phenotype sampling. The range of the trait Y is divided into segments, and individuals from each segment are sampled with different probabilities. These probabilities can be set so that we sample for example only from the extremes. Additionally, Zhou et al. [2002] proposed to include a random sample of size n_0 . Motivated by this design, we consider a design where we first select a random sample of size n_0 to be genotyped, and thereafter sample n_e extreme-phenotype individuals. We assume an EPS-full sampling design so that non-genetic information is available in the full sample. We consider a full data set of size N , and a genotyped sample of size $n_0 + n_e = n$. We consider different sample sizes for the genotyped sample; n ranging from 1000 to 5000, and for each n we consider different sizes of the random and the extreme sample (n_0 and n_e). The parameter values were set as in Table 1 but with $\sigma = 8$ so that the power in the full sample was approximately 80% at the 5% significance level. We simulated $R = 10\,000$ data sets. We used the EPS-full likelihood method to test $H_0 : \beta_g = 0$ against $H_1 : \beta_g \neq 0$ in each sample. The results of the simulation study is presented in Figure 3. We observe that the design with $n_0 = n$ and $n_e = 0$ (a random sample of size n) performs poorly, while all sampling methods that combine extreme and random samples have almost the same power as the EPS-full sample ($n_0 = 0$ and $n_e = n$).

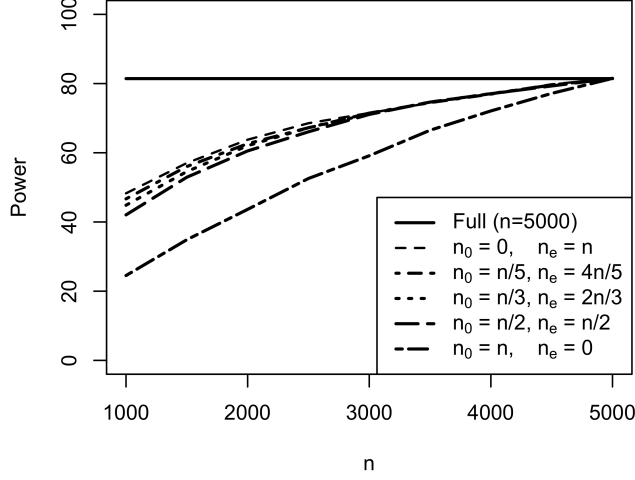


Figure 3: Estimated power to detect a non-null genetic effect ($H_0 : \beta_g = 0$) for outcome-dependent sampling in simulation model (5) for increasing number of genotyped individuals ($n = n_0 + n_e$, where n_0 is the size of the random sample and n_e is the size of the extreme sample), compared to the power for the full sample where $n = N = 5000$.

6 Discussion

We have here considered a sampling design for genetic association studies that has been proposed to increase power of genetic association studies with limited sample sizes. Under the EPS design, individuals with an extreme phenotype are selected for genotyping. If a true association exists between the trait and a genetic variant, the extreme sample will be enriched with homozygous individuals, i.e. individuals with none or two copies of the minor-allele (assuming an additive effect). We have presented relevant statistical methods for this design; some methods are currently used (EPS-only binary) and some methods have to a lesser extent been used in practice (EPS-only continuous and EPS-full). The EPS-only binary method is a valid choice for extreme phenotype samples, and it is also simple to use. However, we have shown that the dichotomization of the continuous trait eliminates any potential gain in power due to extreme sampling, as compared to random sampling. The EPS-only continuous method is a likelihood method that takes into account the continuous probability distribution of the extremes. For this method, we have shown how to obtain parameter estimates and perform hypothesis tests for parameters of a linear regression model that is assumed to hold in the full population. Using the EPS-only likelihood method, we have seen that extreme samples can be more powerful than random samples. In some studies non-genetic variables will be known for the full population (or large sample) while only the extreme phenotype individuals are genotyped. We then have a missing covariate sample where the missing-mechanism is MAR. The EPS-full likelihood was derived based on this principle. We have shown through simulations and real data applications that the power of the EPS-full design can be similar to analyzing the full population. The EPS-full sample can also be analyzed using a multiple imputation approach. In our simulated data the EPS-full MI method had low power, while the then correctly specified EPS-full likelihood model performed significantly better. In the application to real data the multiple imputation method performed similarly to the likelihood method. Multiple imputation could be a useful method for analysis of extreme sampling data, but methodological improvements towards computational efficiency should be considered.

We have extended current statistical methods for EPS-data to include both genetic and non-genetic (environmental) covariates in order to comply with practical situations. Derivation of the score test for the EPS-only and EPS-full design is algebraically tedious and details are therefore only

given in the appendix. We consider the computational efficiency of the score test to be important for genome-wide analysis where thousands of tests are performed. We also derived likelihood methods for estimating and testing gene-environment interaction effects. This was based on relevance for genetic association studies, as previously experienced by us [Bjørnland et al., 2016]. We showed that extreme phenotype sampling can be more powerful than random sampling when testing for gene-environment interaction effects. For specific gene-environment interaction situations we also showed that the extreme exposure design can be more powerful than both random and extreme phenotype sampling.

Some limitations of extreme sampling for genetic association studies also need to be discussed. First of all, extreme phenotype sampling is a complicated procedure compared to random sampling and requires sophisticated statistical methods. Second, we base our methods on the assumption of normality. This assumption can be difficult to check when one only has extreme data. Thirdly, the data could be difficult to include in larger studies (consortia, meta-analysis) due to the non-standard design. Furthermore, the issue of confounding (for example due to population stratification) is of relevance and the EPS-full method is particularly sensitive to unobserved confounding effects (see Appendix C for an illustration of this issue in the HUNT data). Lastly, for genetic association studies and GWAS in particular there are several quality control procedures that are done prior to data analysis. In an extreme sample it is not clear that all QC methods apply directly. Therefore, we advocate to supplement the extreme sample with a random sample so that the random sample can be used for checking model assumptions, confounding, etc. We showed in the last Section of this paper that a mixture of a random sample and an EPS-full sample can be almost as powerful as an EPS-full sample.

In conclusion, we have presented methods for extreme phenotype sampling and shown that the design *can* give better power than random sampling in genetic association studies where the sample size for genotyping is limited. However, only one method (EPS-full) was here found to be clearly better, while other methods were slightly better (EPS-only continuous) or often worse (EPS-only binary) than random sampling. We have extended methods towards testing and modeling of gene-environment interaction effects and shown that also for this purpose, extreme phenotype sampling can be useful. If the appropriate statistical methods are used, it is possible to achieve significantly improved power in genetic association studies with continuous traits by genotyping extreme-phenotype individuals instead of a random sample.

Software

Software is available as an R-package at <https://github.com/theabjorn/extremesampling>.

Acknowledgments

The Nord-Trøndelag Health Study (the HUNT study) is collaboration between the HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology), the Nord-Trøndelag County Council, the Central Norway Health Authority, and the Norwegian Institute of Public Health.

Conflict of interest

None declared.

A The EPS-only likelihood and tests

The linear regression model we have studied is given by

$$Y_i = \alpha + \mathbf{x}_{ei}^T \boldsymbol{\beta}_e + \mathbf{x}_{gi}^T \boldsymbol{\beta}_g + (\mathbf{x}_{ei} \mathbf{x}_{gi})^T \boldsymbol{\beta}_{eg} + \varepsilon_i, \quad \varepsilon_i \text{ i.i.d } \mathcal{N}(0, \sigma^2). \quad (8)$$

Here, Y_i is the phenotype of individual i , \mathbf{x}_{ei} is a vector of environmental (non-genetic) covariates, while \mathbf{x}_{gi} is a vector of genetic covariates (SNP genotypes for a set of SNPs). The term $\mathbf{x}_{ei} \mathbf{x}_{gi}$ is a vector of interactions between relevant environmental and genetic covariates, e.g. $\mathbf{x}_{ei} \mathbf{x}_{gi} = (x_{eij} x_{gik}, x_{eij} x_{gil})$ for some $j \in \{1, \dots, d\}$ and $k, l \in \{1, \dots, m\}$, $k \neq l$. For the EPS-only design, the observations $(y_i, \mathbf{x}_{ei}, \mathbf{x}_{gi})$ are available for all individuals $i \in \mathcal{C}$ for which $y_i < c_l$ or $y_i > c_u$. Let Y_{ci} denote a random variable from the extremes of the distribution of Y_i . We derive the probability distribution of Y_{ci} for known \mathbf{x}_{ei} and \mathbf{x}_{gi} . For simplicity, we denote $F_{Y_c}(y; \mathbf{x}_e, \mathbf{x}_g, \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg}, \sigma, c_l, c_u)$ by $F_{Y_c}(y)$, etc. The distribution of Y_c is then;

$$\begin{aligned} F_{Y_c}(y) &= P(Y_c \leq y) \\ &= P(Y \leq y | (Y < c_l) \cup (Y > c_u)) \\ &= \frac{P(Y \leq y \cap ((Y < c_l) \cup (Y > c_u)))}{P((Y < c_l) \cup (Y > c_u))} \\ &= \frac{P(((Y \leq y) \cap (Y < c_l)) \cup ((Y \leq y) \cap (Y > c_u)))}{P((Y < c_l) \cup (Y > c_u))} \\ &= \frac{P((Y \leq y) \cap (Y < c_l)) + P((Y \leq y) \cap (Y > c_u)) - P((Y \leq y) \cap (Y < c_l) \cap (Y > c_u))}{P((Y < c_l) \cup (Y > c_u))} \\ &= \frac{P((Y \leq y) \cap (Y < c_l)) + P((Y \leq y) \cap (Y > c_u))}{P((Y < c_l) \cup (Y > c_u))} \\ &= \begin{cases} \frac{P(Y \leq y)}{P((Y < c_l) \cup (Y > c_u))} & \text{if } y < c_l \\ 0 & \text{if } c_l < y < c_u \\ \frac{P(Y < c_l) + P(Y \leq y) - P(Y \leq c_u)}{P((Y < c_l) \cup (Y > c_u))} & \text{if } y > c_u \end{cases} \end{aligned}$$

We have defined Y to be normally distributed with mean $\mu(\mathbf{x}_e, \mathbf{x}_g; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg}, \sigma) = \alpha + \mathbf{x}_e^T \boldsymbol{\beta}_e + \mathbf{x}_g^T \boldsymbol{\beta}_g + (\mathbf{x}_e \mathbf{x}_g)^T \boldsymbol{\beta}_{eg}$ and variance σ^2 . We write

$$P(Y \leq y) = \Phi\left(\frac{y - \mu(\mathbf{x}_e, \mathbf{x}_g; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right)$$

where $\Phi()$ represents the cumulative probability distribution of the standard normal distribution. Let $\phi()$ denote the density function of the standard normal distribution. The probability density function for Y_c is given by

$$\begin{aligned} f_{Y_c}(y) &= \frac{\partial}{\partial y} F_{Y_c}(y) \\ &= \begin{cases} \frac{\frac{1}{\sigma} \phi\left(\frac{y - \mu(\mathbf{x}_e, \mathbf{x}_g; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right)}{1 - \Phi\left(\frac{c_u - \mu(\mathbf{x}_e, \mathbf{x}_g; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right) + \Phi\left(\frac{c_l - \mu(\mathbf{x}_e, \mathbf{x}_g; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right)} & \text{if } y < c_l \text{ or } y > c_u, \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The likelihood for the continuous EPS-only design is then

$$L = \prod_{i \in \mathcal{C}} \frac{\frac{1}{\sigma} \phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right)}{1 - \Phi\left(\frac{c_u - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right) + \Phi\left(\frac{c_l - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right)},$$

and the log-likelihood is

$$l = \sum_{i \in \mathcal{C}} \log\left(\frac{1}{\sigma} \phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \boldsymbol{\beta}_e, \boldsymbol{\beta}_g, \boldsymbol{\beta}_{eg})}{\sigma}\right)\right)$$

$$- \sum_{i \in \mathcal{C}} \log \left(1 - \Phi \left(\frac{c_u - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma} \right) + \Phi \left(\frac{c_l - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma} \right) \right)$$

For simpler notation, define $\Phi_{u,i} = \Phi \left(\frac{c_u - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma} \right)$, and similarly for $\Phi_{l,i}$. Then

$$l \propto -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i \in \mathcal{C}} (y_i - \alpha - \mathbf{x}_{ei}^T \beta_e - \mathbf{x}_{gi}^T \beta_g - (\mathbf{x}_{ei} \mathbf{x}_{gi})^T \beta_{eg})^2 - \sum_{i \in \mathcal{C}} \log(1 - \Phi_{u,i} + \Phi_{l,i})$$

For hypothesis testing it is not necessary to distinguish between non-genetic covariates, genetic covariates and second-order interactions and we rewrite the linear model as

$$Y_i = \alpha + \mathbf{x}_{0i}^T \beta_0 + \mathbf{x}_{1i}^T \beta_1 + \varepsilon_i,$$

where under the null hypothesis $\beta_0 = 0$, while β_1 , α and σ are nuisance parameters. Thus \mathbf{x}_0 is a vector of covariates that we want to test for association with Y (e.g. \mathbf{x}_g), while \mathbf{x}_1 is a vector of all other covariates (e.g. \mathbf{x}_e). The log-likelihood can then be written as

$$l \propto -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i \in \mathcal{C}} (y_i - \alpha - \mathbf{x}_{0i}^T \beta_0 - \mathbf{x}_{1i}^T \beta_1)^2 - \sum_{i \in \mathcal{C}} \log(1 - \Phi_{u,i} + \Phi_{l,i}). \quad (9)$$

A.1 The score test

We derive the score test statistic for the two-sided null hypothesis $H_0 : \beta_0 = 0$ vs $H_1 : \beta_0 \neq 0$. Let θ_0 denote all the parameters in the null model, α , β_1 and σ , and let $\hat{\theta}_0$ denote maximum likelihood estimators under the null hypothesis. The score vector is given by the first derivative of the log-likelihood with respect to β_0 , evaluated in $\beta_0 = 0$ and $\hat{\theta}_0$;

$$S = \frac{\partial l}{\partial \beta_1}(\hat{\theta}_0, \beta_0 = 0).$$

The variance of the score vector is given by

$$\Sigma = I_{\beta_0, \beta_0}(\hat{\theta}_0, \beta_0 = 0) - I_{\beta_0, \theta_0}(\hat{\theta}_0, \beta_0 = 0) I_{\theta_0, \theta_0}(\hat{\theta}_0, \beta_0 = 0)^{-1} I_{\theta_0, \beta_0}(\hat{\theta}_0, \beta_0 = 0),$$

where I_{β_0, β_0} is the element of the information matrix corresponding to β_0 , etc. The score test statistic,

$$T = S^T \Sigma^{-1} S$$

is asymptotically χ^2 -distributed under the null hypothesis, with degrees of freedom equal to the number of parameters that we test (i.e. the length of β_0).

In order to derive the score test statistic for the EPS-only likelihood we need the first and second derivatives of the log-likelihood (9). For simpler notation, we write $f_i = y_i - \alpha - \mathbf{x}_{0i}^T \beta_0 - \mathbf{x}_{1i}^T \beta_1$ and we define the following functions (extensions of similar functions defined by Tang [2010]):

$$h_{ij} = \frac{-\phi_{u,i} \cdot \left(\frac{c_u - \mu(\mathbf{x}_{0i}, \mathbf{x}_{1i}; \alpha, \beta_0, \beta_1)}{\sigma} \right)^j + \phi_{l,i} \cdot \left(\frac{c_l - \mu(\mathbf{x}_{0i}, \mathbf{x}_{1i}; \alpha, \beta_0, \beta_1)}{\sigma} \right)^j}{1 - \Phi_{u,i} + \Phi_{l,i}},$$

for $j = 0, 1, 2, 3$, and furthermore

$$\begin{aligned} a_i &= 1 - h_{i1} - h_{i0}^2 \\ b_i &= h_{i0} - h_{i2} - h_{i0} h_{i1} \\ c_i &= -1 + 2h_{i1} - h_{i3} - h_{i1}^2 \\ d_i &= 2 + 2h_{i1} - h_{i3} - h_{i1}^2. \end{aligned}$$

The first derivatives are

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\sigma^2} \sum_{i=1}^n f_i + \frac{1}{\sigma} \sum_{i=1}^n h_{i0},$$

$$\begin{aligned}
\frac{\partial l}{\partial \beta_0} &= \frac{1}{\sigma^2} \sum_{i=1}^n f_i \mathbf{x}_{0i}^T + \frac{1}{\sigma} \sum_{i=1}^n h_{i0} \mathbf{x}_{0i}^T, \\
\frac{\partial l}{\partial \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^n f_i \mathbf{x}_{1i}^T + \frac{1}{\sigma} \sum_{i=1}^n h_{i0} \mathbf{x}_{1i}^T, \\
\frac{\partial l}{\partial \sigma} &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n f_i^2 + \frac{1}{\sigma} \sum_{i=1}^n h_{i1},
\end{aligned}$$

and score vector can now be written as

$$S = \frac{1}{\sigma^2} \sum_{i=1}^n f_i(\hat{\boldsymbol{\theta}}_0, \beta_0 = 0) \mathbf{x}_{0i}^T + \frac{1}{\sigma} \sum_{i=1}^n h_{i0}(\hat{\boldsymbol{\theta}}_0, \beta_0 = 0) \mathbf{x}_{0i}^T.$$

The second derivatives of the log likelihood are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha^2} &= \frac{1}{\sigma^2} \sum_{i=1}^n (-1 + h_{i1} + h_{i0}^2) = -\frac{1}{\sigma^2} \sum_{i=1}^n a_i, \\
\frac{\partial^2 l}{\partial \alpha \partial \beta_0} &= \frac{1}{\sigma^2} \sum_{i=1}^n (-1 + h_{i1} + h_{i0}^2) \mathbf{x}_{0i}^T = -\frac{1}{\sigma^2} \sum_{i=1}^n a_i \mathbf{x}_{0i}^T, \\
\frac{\partial^2 l}{\partial \alpha \partial \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^n (-1 + h_{i1} + h_{i0}^2) \mathbf{x}_{1i}^T = -\frac{1}{\sigma^2} \sum_{i=1}^n a_i \mathbf{x}_{1i}^T, \\
\frac{\partial^2 l}{\partial \alpha \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^n f_i + \frac{1}{\sigma^2} \sum_{i=1}^n (-h_{i0} + h_{2i} + h_{i0} h_{i1}) = -\frac{2}{\sigma^3} \sum_{i=1}^n f_i - \frac{1}{\sigma^2} \sum_{i=1}^n b_i, \\
\frac{\partial^2 l}{\partial \beta_0^T \beta_0} &= \frac{1}{\sigma^2} \sum_{i=1}^n (-1 + h_{i1} + h_{i0}^2) \mathbf{x}_{0i} \mathbf{x}_{0i}^T = -\frac{1}{\sigma^2} \sum_{i=1}^n a_i \mathbf{x}_{0i} \mathbf{x}_{0i}^T, \\
\frac{\partial^2 l}{\partial \beta_0^T \partial \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^n (-1 + h_{i1} + h_{i0}^2) \mathbf{x}_{0i} \mathbf{x}_{1i}^T = -\frac{1}{\sigma^2} \sum_{i=1}^n a_i \mathbf{x}_{0i} \mathbf{x}_{1i}^T, \\
\frac{\partial^2 l}{\partial \beta_1^T \beta_1} &= \frac{1}{\sigma^2} \sum_{i=1}^n (-1 + h_{i1} + h_{i0}^2) \mathbf{x}_{1i} \mathbf{x}_{1i}^T = -\frac{1}{\sigma^2} \sum_{i=1}^n a_i \mathbf{x}_{1i} \mathbf{x}_{1i}^T, \\
\frac{\partial^2 l}{\partial \beta_0 \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^n f_i \mathbf{x}_{0i}^T + \frac{1}{\sigma^2} \sum_{i=1}^n (-h_{i0} + h_{2i} + h_{i0} h_{i1}) \mathbf{x}_{0i}^T = -\frac{2}{\sigma^3} \sum_{i=1}^n f_i \mathbf{x}_{0i}^T - \frac{1}{\sigma^2} \sum_{i=1}^n b_i \mathbf{x}_{0i}^T, \\
\frac{\partial^2 l}{\partial \beta_1 \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i=1}^n f_i \mathbf{x}_{1i}^T + \frac{1}{\sigma^2} \sum_{i=1}^n (-h_{i0} + h_{2i} + h_{i0} h_{i1}) \mathbf{x}_{1i}^T = -\frac{2}{\sigma^3} \sum_{i=1}^n f_i \mathbf{x}_{1i}^T - \frac{1}{\sigma^2} \sum_{i=1}^n b_i \mathbf{x}_{1i}^T, \\
\frac{\partial^2 l}{\partial \sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^n f_i^2 + \frac{1}{\sigma^2} \sum_{i=1}^n (-2h_{i1} - h_{i2} + h_{i1}^2) = -\frac{3}{\sigma^4} \sum_{i=1}^n f_i^2 - \frac{1}{\sigma^2} \sum_{i=1}^n c_i.
\end{aligned}$$

For simplicity, let $a_i(\boldsymbol{\theta}_0, \beta_0 = 0)$ be denoted by a_i^0 , etc. Then,

$$\begin{aligned}
I_{\beta_0, \beta_0}(\boldsymbol{\theta}_0, \beta_0 = 0) &= -\mathbb{E} \left(\frac{\partial^2 l}{\partial \beta_0^T \partial \beta_0} \right) \Big|_{(\boldsymbol{\theta}_0, \beta_0 = 0)} = \frac{1}{\sigma^2} \sum_{i=1}^n a_i^0 \mathbf{x}_{0i} \mathbf{x}_{0i}^T, \\
I_{\boldsymbol{\theta}_0, \beta_0}(\boldsymbol{\theta}_0, \beta_0 = 0) &= -\begin{bmatrix} \mathbb{E} \left(\frac{\partial^2 l}{\partial \alpha \partial \beta_0} \right) \\ \mathbb{E} \left(\frac{\partial^2 l}{\partial \beta_1^T \partial \beta_0} \right) \\ \mathbb{E} \left(\frac{\partial^2 l}{\partial \sigma \partial \beta_0} \right) \end{bmatrix} \Big|_{(\boldsymbol{\theta}_0, \beta_0 = 0)} = \frac{1}{\sigma^2} \begin{bmatrix} \sum_{i=1}^n a_i^0 \mathbf{x}_{0i} \mathbf{x}_{0i}^T \\ \sum_{i=1}^n a_i^0 \mathbf{x}_{1i} \mathbf{x}_{0i}^T \\ \sum_{i=1}^n b_i^0 \mathbf{x}_{0i} \mathbf{x}_{0i}^T \end{bmatrix},
\end{aligned}$$

$I_{\beta_0, \theta_0}(\theta_0, \beta_0 = 0) = I_{\theta_0, \beta_0}(\theta_0, \beta_0 = 0)^T$, and

$$\begin{aligned} I_{\theta_0, \theta_0}(\theta_0, \beta_0 = 0) &= - \begin{bmatrix} E\left(\frac{\partial^2 l}{\partial \alpha^2}\right) & E\left(\frac{\partial^2 l}{\partial \alpha \partial \beta_1}\right) & E\left(\frac{\partial^2 l}{\partial \alpha \partial \sigma}\right) \\ E\left(\frac{\partial^2 l}{\partial \alpha \partial \beta_1}\right)^T & E\left(\frac{\partial^2 l}{\partial \beta_1^T \partial \beta_1}\right) & E\left(\frac{\partial^2 l}{\partial \beta_1 \partial \sigma}\right)^T \\ E\left(\frac{\partial^2 l}{\partial \alpha \partial \sigma}\right) & E\left(\frac{\partial^2 l}{\partial \beta_1 \partial \sigma}\right) & E\left(\frac{\partial^2 l}{\partial \sigma^2}\right) \end{bmatrix}_{(\theta_0, \beta_0=0)} \\ &= \frac{1}{\sigma^2} \begin{bmatrix} \sum_{i=1}^n a_i^0 & \sum_{i=1}^n a_i^0 \mathbf{x}_{1i}^T & \sum_{i=1}^n b_i^0 \\ \sum_{i=1}^n a_i^0 \mathbf{x}_{1i} & \sum_{i=1}^n a_i^0 \mathbf{x}_{1i} \mathbf{x}_{1i}^T & \sum_{i=1}^n b_i^0 \mathbf{x}_{1i} \\ \sum_{i=1}^n b_i^0 & \sum_{i=1}^n b_i^0 \mathbf{x}_{1i}^T & \sum_{i=1}^n d_i^0 \end{bmatrix}. \end{aligned}$$

In the last expression we have used the fact that $E(f_i) = 0$ and $E(f_i^2) = \sigma^2$, for example to derive $-E\left(\frac{\partial^2 l}{\partial \sigma^2}\right) = -E\left(-\frac{3}{\sigma^4} \sum_{i=1}^n f_i^2 - \frac{1}{\sigma^2} \sum_{i=1}^n c_i^0\right) = \frac{1}{\sigma^2} \sum_{i=1}^n 3 + c_i^0 = \frac{1}{\sigma^2} \sum_{i=1}^n d_i^0$. The variance Σ of the score vector can be obtained by evaluating these expression in $\hat{\theta}_0$, the maximum likelihood estimates under the null.

For the special case when no other covariates than \mathbf{x}_0 are included in the model, then h_{ij} , a_i , b_i , c_i and d_i evaluated in $\mathbf{x}_g = 0$ does not depend on the index i . Furthermore, the maximum likelihood estimators $\hat{\alpha}$ and $\hat{\sigma}^2$ under the null hypothesis can be found by solving

$$\begin{aligned} 0 &= \frac{1}{\sigma^2} \sum_{i=1}^n (y_i - \alpha) + \frac{1}{\sigma} n h_0 \\ 0 &= -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^n (y_i - \alpha)^2 + \frac{1}{\sigma} n h_1, \end{aligned}$$

which yields

$$\begin{aligned} \hat{\sigma}^2 &= \frac{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}{a}, \\ \hat{\alpha} &= \bar{y} + \hat{\sigma} h_0. \end{aligned}$$

The score vector is given by

$$S = \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n (y_i - \hat{\alpha}) \mathbf{x}_{0i}^T + \frac{1}{\hat{\sigma}} h_0 \sum_{i=1}^n \mathbf{x}_{0i}^T = a \frac{\sum_{i=1}^n (y_i - \bar{y}) \mathbf{x}_{0i}^T}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} = a S^*,$$

where S^* is the score vector derived from a linear regression model for a random sample. Furthermore, the variance of the score vector becomes

$$\begin{aligned} \Sigma &= \frac{a}{\hat{\sigma}^2} \sum_{i=1}^n \mathbf{x}_{0i} \mathbf{x}_{0i}^T - \frac{1}{\hat{\sigma}^2} \sum_{i=1}^n \mathbf{x}_{0i} \begin{bmatrix} a & a \end{bmatrix} \begin{bmatrix} na & nb \\ nb & nd \end{bmatrix}^{-1} \sum_{i=1}^n \mathbf{x}_{0i}^T \begin{bmatrix} a \\ a \end{bmatrix} \\ &= \frac{a^2}{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2} \sum_{i=1}^n (\mathbf{x}_{0i} - \bar{\mathbf{x}}_0)(\mathbf{x}_{0i} - \bar{\mathbf{x}}_0)^T = a^2 \Sigma^*, \end{aligned}$$

where Σ^* is the variance of the score vector derived from a linear regression model for a random sample. Then $T = S^T \Sigma^{-1} S = (aS^*)^T (a^2 \Sigma^*)^{-1} a S^* = (S^*)^T (\Sigma^*)^{-1} S^* = T^*$, i.e. equivalent to the score test statistic derived under the assumption that the we have a random sample and not an EPS-only sample. This confirms the finding of Tang [2010].

B The EPS-full likelihood and tests

We consider a regression model as in (8). In the EPS-full design we have observed the variables $(Y_i, \mathbf{X}_{ei}, \mathbf{X}_{gi})$ for all individuals $i \in \mathcal{C}$ and (Y_i, \mathbf{X}_{ei}) for all individuals $i \notin \mathcal{C}$. The genetic covariate \mathbf{X}_g is missing at random (MAR) and we therefore derive the likelihood that ignores the missing-mechanism [Little and Rubin, 2002]. The joint density of Y_i , \mathbf{X}_{ei} and \mathbf{X}_{gi} is $f_{Y_i, \mathbf{X}_{ei}, \mathbf{X}_{gi}}(y_i, \mathbf{x}_{ei}, \mathbf{x}_{gi})$.

The marginal distribution of the observed data is defined by integrating out missing data. In our case, we have a discrete missing covariate \mathbf{X}_{gi} . For $i \notin \mathcal{C}$ the variable \mathbf{X}_{gi} is missing such that the marginal distribution of the observed variables is $f_{Y_i, \mathbf{X}_{ei}}(y_i, \mathbf{x}_{ei}; \theta) = \sum_{\mathbf{x}_g} f_{Y_i, \mathbf{X}_{ei}, \mathbf{X}_{gi}}(y_i, \mathbf{x}_{ei}, \mathbf{x}_g)$. The likelihood that *ignores* the missing-mechanism is then

$$\begin{aligned} L_{ign} &= \prod_{i \in \mathcal{C}} f_{Y_i, \mathbf{X}_{ei}, \mathbf{X}_{gi}}(y_i, \mathbf{x}_{ei}, \mathbf{x}_{gi}) \prod_{i \notin \mathcal{C}} \sum_{\mathbf{x}_g} f_{Y_i, \mathbf{X}_{ei}, \mathbf{X}_{gi}}(y_i, \mathbf{x}_{ei}, \mathbf{x}_g) \\ &= \prod_{i \in \mathcal{C}} f_{Y_i | \mathbf{X}_{ei}=\mathbf{x}_{ei}, \mathbf{X}_{gi}=\mathbf{x}_{gi}}(y_i) f_{\mathbf{X}_{gi} | \mathbf{X}_{ei}=\mathbf{x}_{ei}}(\mathbf{x}_{gi}) f_{\mathbf{X}_{ei}}(\mathbf{x}_{ei}) \\ &\quad \prod_{i \notin \mathcal{C}} \sum_{\mathbf{x}_g} f_{Y_i | \mathbf{X}_{ei}=\mathbf{x}_{ei}, \mathbf{X}_{gi}=\mathbf{x}_g}(y_i) f_{\mathbf{X}_{gi} | \mathbf{X}_{ei}=\mathbf{x}_{ei}}(\mathbf{x}_g) f_{\mathbf{X}_{ei}}(\mathbf{x}_{ei}). \end{aligned}$$

We write $f_{Y_i | \mathbf{X}_{ei}=\mathbf{x}_{ei}, \mathbf{X}_{gi}=\mathbf{x}_{gi}}(y_i) = \frac{1}{\sigma} \phi_i$, where $\phi_i = \phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_{gi}; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right)$ and $\phi(\cdot)$ is the standard normal density function. For individuals $i \notin \mathcal{C}$, we write $\phi_i(\mathbf{x}_g)$ to denote $\phi\left(\frac{y_i - \mu(\mathbf{x}_{ei}, \mathbf{x}_g; \alpha, \beta_e, \beta_g, \beta_{eg})}{\sigma}\right)$. We assume that the distribution of \mathbf{X}_g depends on a subset of \mathbf{X}_e , denoted \mathbf{X} , and that the sample space of \mathbf{X} is discrete and of size J . Furthermore, we assume that $f_{\mathbf{X}_{gi} | \mathbf{X}_i=\mathbf{x}_j}(\mathbf{x}_g) = f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_g)$, i.e. that the distribution of genotypes is the same for all individuals with equal value of \mathbf{X} . We assume that the sample space of \mathbf{X}_g is discrete and of size K . The likelihood is then

$$L \propto \prod_{i \in \mathcal{C}} \frac{1}{\sigma} \phi_i \sum_{j=1}^J f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gi}) I(\mathbf{x}_i = \mathbf{x}_j) \prod_{i \notin \mathcal{C}} \sum_{k=1}^K \frac{1}{\sigma} \phi_i(\mathbf{x}_{gk}) \sum_{j=1}^J f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) I(\mathbf{x}_i = \mathbf{x}_j)$$

The log-likelihood for the EPS-full model can then be written as

$$\begin{aligned} l &\propto -N \log(\sigma) + \sum_{i \in \mathcal{C}} \left(\log(\phi_i) + \sum_{j=1}^J \log(f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gi})) I(\mathbf{x}_i = \mathbf{x}_j) \right) \\ &\quad + \sum_{i \notin \mathcal{C}} \log \left(\sum_{j=1}^J \sum_{k=1}^K \phi_i(\mathbf{x}_{gk}) f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) I(\mathbf{x}_i = \mathbf{x}_j) \right) \end{aligned}$$

B.1 The score test

We derive the score test for the two-sided null hypothesis $H_0 : \beta_g = 0$ in the model $y = \alpha + \mathbf{x}_e^T \beta_e + \mathbf{x}_g^T \beta_g + \varepsilon$. In this case, the model under the null is simply a linear regression model that can be fitted with standard methods due to the complete sampling of (y_i, \mathbf{x}_{ei}) . In our work with common genetic variants, we consider genetic variants that can take three possible values (0, 1 or 2). To obtain a more general result we consider a discrete sample space of \mathbf{X}_g of size K . We then have $P(\mathbf{X}_g = \mathbf{x}_{gk} | \mathbf{X} = \mathbf{x}_j) = p_{jk}$ for $k = 1, \dots, K-1$ and $P(\mathbf{X}_g = \mathbf{x}_{gK} | \mathbf{X} = \mathbf{x}_j) = 1 - \sum_{k=1}^{K-1} p_{jk}$, for $j = 1, \dots, J$. Let θ_0 denote all parameters in the null model (α, β_e, σ , and $p_{jk}, j = 1, \dots, J, k = 1, \dots, K-1$). For simplified notation, we write $f_i = y_i - \alpha - \mathbf{x}_{ei}^T \beta_e - \mathbf{x}_{gi}^T \beta_g$ and $f_i(\mathbf{x}_g) = y_i - \alpha - \mathbf{x}_{ei}^T \beta_e - \mathbf{x}_g^T \beta_g$. We write $\phi_i(\mathbf{x}_g)$ to denote $\phi(f_i(\mathbf{x}_g)/\sigma)$. We define

$$h_{iab} = \frac{\sum_j \sum_k f_i(\mathbf{x}_{gk})^a \phi_i(\mathbf{x}_{gk}) f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) \mathbf{x}_{gk}^b I(\mathbf{x}_i = \mathbf{x}_j)}{\sum_j \sum_k \phi_i(\mathbf{x}_{gk}) f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) I(\mathbf{x}_i = \mathbf{x}_j)},$$

where $\mathbf{x}_{gk}^0 = 1$, $\mathbf{x}_{gk}^1 = \mathbf{x}_{gk}^T$ and $\mathbf{x}_{gk}^2 = \mathbf{x}_{gk} \mathbf{x}_{gk}^T$. We also define

$$h_{iab}^{(j'k')} = \frac{\left(f_i(\mathbf{x}_{gk'})^a \phi_i(\mathbf{x}_{gk'}) \mathbf{x}_{gk'}^b - f_i(\mathbf{x}_{gK})^a \phi_i(\mathbf{x}_{gK}) \mathbf{x}_{gK}^b \right) I(\mathbf{x}_i = \mathbf{x}_{j'})}{\sum_j \sum_k \phi_i(\mathbf{x}_{gk}) f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) I(\mathbf{x}_i = \mathbf{x}_j)}.$$

Note that $h_{i00} = 1$. Furthermore, we will use that

$$h_{iab}(\beta_g = 0) = \frac{\sum_j \sum_k f_i(\beta_g = 0)^a \phi_i(\beta_g = 0) f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) \mathbf{x}_{gk}^b I(\mathbf{x}_i = \mathbf{x}_j)}{\sum_j \sum_k \phi_i(\beta_g = 0) f_{\mathbf{X}_g | \mathbf{X}=\mathbf{x}_j}(\mathbf{x}_{gk}) I(\mathbf{x}_i = \mathbf{x}_j)}$$

$$\begin{aligned}
&= \sum_j f_i(\beta_g = 0)^a \left(\sum_k f_{\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j}(\mathbf{x}_{gk}) \mathbf{x}_{gk}^b \right) I(\mathbf{x}_i = \mathbf{x}_j) \\
&= \sum_j f_i(\beta_g = 0)^a E(\mathbf{X}_g^b | \mathbf{X} = \mathbf{x}_j) I(\mathbf{x}_i = \mathbf{x}_j)
\end{aligned}$$

Let n_{jk} denote the number of individuals $i \in \mathcal{C}$ for which $\mathbf{x}_i = \mathbf{x}_j$ and $\mathbf{x}_{gi} = \mathbf{x}_{gk}$.

The first derivatives of the log-likelihood are

$$\begin{aligned}
\frac{\partial l}{\partial \alpha} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} f_i + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i10}, \\
\frac{\partial l}{\partial \beta_e} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} f_i \mathbf{x}_{ei}^T + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i10} \mathbf{x}_{ei}^T, \\
\frac{\partial l}{\partial \beta_g} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} f_i \mathbf{x}_{gi}^T + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i11}, \\
\frac{\partial l}{\partial \sigma} &= -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{i \in \mathcal{C}} f_i^2 + \frac{1}{\sigma^3} \sum_{i \notin \mathcal{C}} h_{i20}, \\
\frac{\partial l}{\partial p_{j'k'}} &= \frac{n_{j'k'}}{p_{j'k'}} - \frac{n_{j'K}}{1 - \sum_{k=1}^{K-1} p_{j'k}} + \sum_{i \notin \mathcal{C}} h_{i00}^{j'k'}, \text{ for } j' = 1, \dots, J, k' = 1, \dots, K-1,
\end{aligned}$$

and the score vector is then given by

$$\begin{aligned}
S &= \frac{\partial l}{\partial \beta_g}(\hat{\theta}_0, \beta_g = 0) \\
&= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} f_i(\hat{\theta}_0, \beta_g = 0) \mathbf{x}_{gi}^T + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i11}(\hat{\theta}_0, \beta_g = 0) \\
&= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} f_i(\hat{\theta}_0, \beta_g = 0) \mathbf{x}_{gi}^T + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} f_i(\hat{\theta}_0, \beta_g = 0) \sum_{j=1}^J E(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j) I(\mathbf{x}_i = \mathbf{x}_j).
\end{aligned}$$

Below we have calculated the second derivatives of the log-likelihood and evaluated these under the null hypothesis ($\beta_g = 0$). The second derivatives are

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha^2} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} (-1) + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \left(-h_{i00} + \frac{1}{\sigma^2} h_{i20} - \frac{1}{\sigma^2} h_{i10}^2 \right) \\
&= -\frac{N}{\sigma^2} + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} (h_{i20} - h_{i10}^2) \stackrel{\beta_g=0}{=} -\frac{N}{\sigma^2} \\
\frac{\partial^2 l}{\partial \alpha \partial \beta_e} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} (-\mathbf{x}_{ei}^T) + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \left(-h_{i00} \mathbf{x}_{ei}^T + \frac{1}{\sigma^2} h_{i20} \mathbf{x}_{ei}^T - \frac{1}{\sigma^2} h_{i10}^2 \mathbf{x}_{ei}^T \right) \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^N \mathbf{x}_{ei}^T + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} (h_{i20} - h_{i10}^2) \mathbf{x}_{ei}^T \stackrel{\beta_g=0}{=} -\frac{1}{\sigma^2} \sum_{i=1}^N \mathbf{x}_{ei}^T \\
\frac{\partial^2 l}{\partial \alpha \partial \beta_g} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} (-\mathbf{x}_{gi}^T) + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \left(-h_{i01} + \frac{1}{\sigma^2} h_{i21} - \frac{1}{\sigma^2} h_{i10} h_{i11} \right) \\
&= -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{gi}^T + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} -h_{i01} + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} (h_{i21} - h_{i10} h_{i11}) \\
&\stackrel{\beta_g=0}{=} -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{gi}^T - \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \sum_{j=1}^J E(\mathbf{X}_g | \mathbf{X}_e = \mathbf{x}_{ej})^T I(\mathbf{x}_i = \mathbf{x}_j) \\
\frac{\partial^2 l}{\partial \alpha \partial \sigma} &= -\frac{2}{\sigma^3} \sum_{i \in \mathcal{C}} f_i - 2 \frac{1}{\sigma^3} \sum_{i \notin \mathcal{C}} h_{i10} + \frac{1}{\sigma^5} \sum_{i \notin \mathcal{C}} \left(\frac{1}{\sigma^3} h_{i30} - h_{i20} h_{i10} \right) \stackrel{\beta_g=0}{=} -\frac{2}{\sigma^3} \sum_{i=1}^N f_i(0)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \alpha \partial p_{j'k'}} &= \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \left(h_{i10}^{(j'k')} - h_{i00}^{(j'k')} h_{i10} \right) \stackrel{\beta_g=0}{=} 0 \\
\frac{\partial^2 l}{\partial \beta_e^T \beta_e} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} (-\mathbf{x}_{ei} \mathbf{x}_{ei}^T) + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \mathbf{x}_{ei} \left(-h_{i00} \mathbf{x}_{ei}^T + \frac{1}{\sigma^2} h_{i20} \mathbf{x}_{ei}^T - \frac{1}{\sigma^2} h_{i10}^2 \mathbf{x}_{ei}^T \right) \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^N \mathbf{x}_{ei} \mathbf{x}_{ei}^T + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} (h_{i20} - h_{i10}^2) \mathbf{x}_{ei} \mathbf{x}_{ei}^T \stackrel{\beta_g=0}{=} -\frac{1}{\sigma^2} \sum_{i=1}^N \mathbf{x}_{ei} \mathbf{x}_{ei}^T \\
\frac{\partial^2 l}{\partial \beta_e^T \beta_g} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} (-\mathbf{x}_{ei} \mathbf{x}_{gi}^T) + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \mathbf{x}_{ei} \left(-h_{i01} + \frac{1}{\sigma^2} h_{i21} - \frac{1}{\sigma^2} h_{i10} h_{i11} \right) \\
&= -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{ei} \mathbf{x}_{gi}^T - \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \mathbf{x}_{ei} h_{i01} + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} \mathbf{x}_{ei} (h_{i21} - h_{i10} h_{i11}) \\
&\stackrel{\beta_g=0}{=} -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{ei} \mathbf{x}_{gi}^T - \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \mathbf{x}_{ei} \sum_{j=1}^J E(\mathbf{X}_g | \mathbf{X}_e = \mathbf{x}_{ej})^T I(\mathbf{x}_i = \mathbf{x}_j) \\
\frac{\partial^2 l}{\partial \beta_e^T \sigma} &= -\frac{2}{\sigma^3} \sum_{i \in \mathcal{C}} f_i \mathbf{x}_{ei}^T - \frac{2}{\sigma^3} \sum_{i \notin \mathcal{C}} h_{i10} \mathbf{x}_{ei}^T + \frac{1}{\sigma^5} \sum_{i \notin \mathcal{C}} (h_{i30} \mathbf{x}_{ei}^T - h_{i20} h_{i10} \mathbf{x}_{ei}^T) \\
&\stackrel{\beta_g=0}{=} -\frac{2}{\sigma^3} \sum_{i=1}^N f_i(0) \mathbf{x}_{ei}^T \\
\frac{\partial^2 l}{\partial \beta_e p_{j'k'}} &= \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i10}^{j'k'} \mathbf{x}_{ei}^T - h_{i00}^{j'k'} h_{i10} \mathbf{x}_{ei}^T \stackrel{\beta_g=0}{=} \mathbf{0}_{1 \times |\mathbf{x}_e|} \\
\frac{\partial^2 l}{\partial \beta_g^T \beta_g} &= \frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} (-\mathbf{x}_{gi} \mathbf{x}_{gi}^T) + \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \left(-h_{i02} + \frac{1}{\sigma^2} h_{i22} - \frac{1}{\sigma^2} h_{i11}^T h_{i11} \right) \\
&= -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{gi} \mathbf{x}_{gi}^T - \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i02} + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} (h_{i22} - h_{i11}^T h_{i11}) \\
&\stackrel{\beta_g=0}{=} -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{gi} \mathbf{x}_{gi}^T - \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \sum_{j=1}^J E(\mathbf{X}_g \mathbf{X}_g^T | \mathbf{X} = \mathbf{x}_j) I(\mathbf{x}_i = \mathbf{x}_j) \\
&\quad + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} f_i(0)^2 \sum_{j=1}^J (E(\mathbf{X}_g \mathbf{X}_g^T | \mathbf{X} = \mathbf{x}_j) - E(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j) E(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j)^T) I(\mathbf{x}_i = \mathbf{x}_j) \\
&= -\frac{1}{\sigma^2} \sum_{i \in \mathcal{C}} \mathbf{x}_{gi} \mathbf{x}_{gi}^T - \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} \sum_{j=1}^J E(\mathbf{X}_g \mathbf{X}_g^T | \mathbf{X} = \mathbf{x}_j) I(\mathbf{x}_i = \mathbf{x}_j) \\
&\quad + \frac{1}{\sigma^4} \sum_{i \notin \mathcal{C}} \sum_{j=1}^J f_i(0)^2 \text{Var}(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j) I(\mathbf{x}_i = \mathbf{x}_j) \\
\frac{\partial^2 l}{\partial \beta_g \sigma} &= -\frac{2}{\sigma^3} \sum_{i \in \mathcal{C}} f_i \mathbf{x}_{gi}^T - \frac{2}{\sigma^3} \sum_{i \notin \mathcal{C}} h_{i11} + \frac{1}{\sigma^5} \sum_{i \notin \mathcal{C}} (h_{i31} - h_{i20} h_{i11}) \\
&\stackrel{\beta_g=0}{=} -\frac{2}{\sigma^3} \sum_{i \in \mathcal{C}} f_i(0) \mathbf{x}_{gi}^T - \frac{2}{\sigma^3} \sum_{i \notin \mathcal{C}} f_i(0) \sum_{j=1}^J E(\mathbf{X}_g | \mathbf{X}_e = \mathbf{x}_{ej})^T I(\mathbf{x}_i = \mathbf{x}_j) \\
\frac{\partial^2 l}{\partial \beta_g p_{j'k'}} &= \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} h_{i11}^{j'k'} - h_{i00}^{j'k'} h_{i11} \stackrel{\beta_g=0}{=} \frac{1}{\sigma^2} \sum_{i \notin \mathcal{C}} f_i(0) (\mathbf{x}_{gk'}^T - \mathbf{x}_{gK}^T) I(\mathbf{x}_i = \mathbf{x}_{j'}) \\
\frac{\partial^2 l}{\partial \sigma^2} &= \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i \in \mathcal{C}} f_i^2 - \frac{3}{\sigma^4} \sum_{i \notin \mathcal{C}} h_{i20} + \frac{1}{\sigma^6} \sum_{i \notin \mathcal{C}} (h_{i40} - h_{i20}^2) \stackrel{\beta_g=0}{=} \frac{N}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^N f_i(0)^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l}{\partial \sigma p_{j'k'}} &= \frac{1}{\sigma^3} \sum_{i \notin \mathcal{C}} h_{i20}^{j'k'} - h_{i00}^{j'k'} h_{i20} \stackrel{\beta_g=0}{=} 0 \\
\frac{\partial^2 l}{\partial p_{j'k'}^2} &= -\frac{n_{j'k'}}{p_{j'k'}^2} + \frac{n_{j'K}}{\left(1 - \sum_{k=1}^{K-1} p_{j'k}\right)^2} - \sum_{i \notin \mathcal{C}} (h_{i00}^{j'k'})^2 \stackrel{\beta_g=0}{=} -\frac{n_{j'k'}}{p_{j'k'}^2} + \frac{n_{j'K}}{\left(1 - \sum_{k=1}^{K-1} p_{j'k}\right)^2} \\
\frac{\partial^2 l}{\partial p_{j'k''} p_{j'k'}} &= \frac{n_{j'K}}{\left(1 - \sum_{k=1}^{K-1} p_{j'k}\right)^2} - \sum_{i \notin \mathcal{C}} h_{i00}^{j'k'} h_{i00}^{j'k''} \stackrel{\beta_g=0}{=} \frac{n_{j'K}}{\left(1 - \sum_{k=1}^{K-1} p_{j'k}\right)^2} \\
\frac{\partial^2 l}{\partial p_{j''k'} p_{j'k'}} &= \frac{\partial^2 l}{\partial p_{j''k''} p_{j'k'}} = 0.
\end{aligned}$$

To simplify notation, let $\mathbf{p}_j = (p_{j1}, \dots, p_{jK-1})$, $j = 1, \dots, J$, and let $\mathbf{p} = (\mathbf{p}_1, \dots, \mathbf{p}_J)$ be a vector of length $J(K-1)$. We then write

$$\frac{\partial^2 l}{\partial \beta_g^T \mathbf{p}} = \begin{bmatrix} \frac{\partial^2 l}{\partial \beta_g^T \mathbf{p}_1} & \cdots & \frac{\partial^2 l}{\partial \beta_g^T \mathbf{p}_J} \end{bmatrix},$$

and similarly for other derivatives with respect to p_{jk} . We let $\boldsymbol{\theta}'_0$ denote the parameters $(\alpha, \beta_e, \sigma)$. We use the observed information matrix (evaluated in $\hat{\boldsymbol{\theta}}_0$ and $\beta_g = 0$) as an estimate for the true information matrix under the null, i.e $\hat{I} = I_o(\hat{\boldsymbol{\theta}}_0, \beta_g = 0)$. An estimate of the variance of S is then given by

$$\hat{\Sigma} = \hat{I}_{\beta_g, \beta_g} - \hat{I}_{\beta_g, \theta_0} \hat{I}_{\theta_0, \theta_0}^{-1} \hat{I}_{\theta_0, \beta_g},$$

where

$$\begin{aligned}
\hat{I}_{\beta_g, \beta_g} &= -\frac{\partial^2 l}{\partial \beta_g^T \beta_g} \Big|_{(\hat{\boldsymbol{\theta}}_0, \beta_g=0)} \\
\hat{I}_{\theta_0, \beta_g} &= -\begin{bmatrix} \frac{\partial^2 l}{\partial \alpha \beta_g} \\ \frac{\partial^2 l}{\partial \beta_e^T \beta_g} \\ \frac{\partial^2 l}{\partial \sigma \beta_g} \\ \frac{\partial^2 l}{\partial \mathbf{p}^T \beta_g} \end{bmatrix}_{(\hat{\boldsymbol{\theta}}_0, \beta_g=0)} = \begin{bmatrix} \hat{I}_{\theta'_0, \beta_g} \\ \hat{I}_{\mathbf{p}, \beta_g} \end{bmatrix}, \\
\hat{I}_{\theta_0, \theta_0} &= -\begin{bmatrix} \frac{\partial^2 l}{\partial \alpha^2} & \frac{\partial^2 l}{\partial \alpha \partial \beta_e} & \frac{\partial^2 l}{\partial \alpha \partial \sigma} & \frac{\partial^2 l}{\partial \alpha \partial \mathbf{p}} \\ \frac{\partial^2 l}{\partial \beta_e^T \partial \alpha} & \frac{\partial^2 l}{\partial \beta_e^T \beta_e} & \frac{\partial^2 l}{\partial \beta_e^T \partial \sigma} & \frac{\partial^2 l}{\partial \beta_e^T \partial \mathbf{p}} \\ \frac{\partial^2 l}{\partial \sigma \partial \alpha} & \frac{\partial^2 l}{\partial \sigma \partial \beta_e} & \frac{\partial^2 l}{\partial \sigma^2} & \frac{\partial^2 l}{\partial \sigma \partial \mathbf{p}} \\ \frac{\partial^2 l}{\partial \mathbf{p}^T \partial \alpha} & \frac{\partial^2 l}{\partial \mathbf{p}^T \partial \beta_e} & \frac{\partial^2 l}{\partial \mathbf{p}^T \partial \sigma} & \frac{\partial^2 l}{\partial \mathbf{p}^T \mathbf{p}} \end{bmatrix}_{(\hat{\boldsymbol{\theta}}_0, \beta_g=0)} = \begin{bmatrix} \hat{I}_{\theta'_0, \theta'_0} & \mathbf{0} \\ \mathbf{0} & \hat{I}_{\mathbf{p}, \mathbf{p}} \end{bmatrix}.
\end{aligned}$$

The variance estimate $\hat{\Sigma}$ can then be written as

$$\hat{\Sigma} = \hat{I}_{\beta_g, \beta_g} - \hat{I}_{\beta_g, \theta'_0} \hat{I}_{\theta'_0, \theta'_0}^{-1} \hat{I}_{\theta'_0, \beta_g} - \hat{I}_{\beta_g, \mathbf{p}} \hat{I}_{\mathbf{p}, \mathbf{p}}^{-1} \hat{I}_{\mathbf{p}, \beta_g},$$

Under the null, the estimates of p_{jk} are simply $\hat{p}_{jk} = n_{jk}/N_j$ where N_j is the number of individuals with $\mathbf{x}_i = \mathbf{x}_j$, and estimates of $\boldsymbol{\theta}'_0$ can be found by fitting a linear regression model to the completely observed data ($Y = \alpha + \mathbf{x}_e^T \beta_e + \varepsilon$) using standard methods. A recent paper by Derkach et al. [2015] presents a further simplified expression for $\hat{\Sigma}$. We present the calculations used to come to their expression here. We have

$$\hat{I}_{\mathbf{p}, \mathbf{p}} = \text{Diag}(A_1, \dots, A_J),$$

where each block matrix A_j , $j = 1, \dots, J$ is a $(K-1) \times (K-1)$ matrix with diagonal elements

$$(A_j)_{kk} = \frac{n_{jk}}{\hat{p}_{jk}^2} - \frac{n_{jK}}{\left(1 - \sum_{k'=1}^{K-1} \hat{p}_{jk'}\right)^2}$$

and off-diagonal elements

$$(A_j)_{kk'} = -\frac{n_{jK}}{\left(1 - \sum_{k=1}^{K-1} \hat{p}_{jk}\right)^2}, \quad k \neq k', \text{ and } k, k' = 1, \dots, K-1.$$

Note that A_j equals the observed information matrix for the multinomially distributed variable $\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j \sim (p_{j1}, \dots, p_{jK}; N_j)$, where $p_{jK} = 1 - \sum_{k=1}^{K-1} p_{jk}$. The inverse of this matrix (A_j^{-1}) is then the estimated covariance matrix for the maximum likelihood estimators of the parameters $p_{j1}, \dots, p_{j(K-1)}$. The MLE for p_{jk} is $\hat{p}_{jk} = N_{jk}/N_j$, where N_{jk} is the number of outcomes of type k of N_j trials. Then,

$$\begin{aligned} (A_j^{-1})_{kk} &= \text{Var}(\hat{p}_{jk}) = \frac{1}{N_j^2} \text{Var}(N_{jk}) = \frac{1}{N_j} p_{jk}(1 - p_{jk}) \\ (A_j^{-1})_{kk'} &= \text{Covar}(\hat{p}_{jk}, \hat{p}_{jk'}) = \frac{1}{N_j^2} \text{Covar}(N_{jk}, N_{jk'}) = \frac{1}{N_j} p_{jk} p_{jk'} \end{aligned}$$

for $k \neq k'$ and $k, k' = 1, \dots, K-1$. Then

$$\begin{aligned} \hat{I}_{\beta_g, \mathbf{P}} \hat{I}_{\mathbf{P}, \mathbf{P}}^{-1} \hat{I}_{\mathbf{P}, \beta_g} &= \left[\begin{array}{ccc} \frac{\partial^2 l}{\partial \beta_g^T \mathbf{P}_1 \beta_g} & \cdots & \frac{\partial^2 l}{\partial \beta_g^T \mathbf{P}_J \beta_g} \end{array} \right]_{(\hat{\theta}_0, \beta_g=0)} \text{Diag}(A_1^{-1}, \dots, A_J^{-1}) \begin{bmatrix} \frac{\partial^2 l}{\partial \mathbf{P}_1^T \beta_g} \\ \vdots \\ \frac{\partial^2 l}{\partial \mathbf{P}_J^T \beta_g} \end{bmatrix}_{(\hat{\theta}_0, \beta_g=0)} \\ &= \sum_{j=1}^J \left(\sum_{i \notin \mathcal{C}} f_i(\hat{\theta}_0, \beta_g=0) I(\mathbf{x}_i = \mathbf{x}_j) \right)^2 \boldsymbol{\alpha}_j^T A_j^{-1} \boldsymbol{\alpha}_j \\ &= \sum_{j=1}^J \left(\sum_{i \notin \mathcal{C}} f_i(\hat{\theta}_0, \beta_g=0) I(\mathbf{x}_i = \mathbf{x}_j) \right)^2 \text{Var}(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j) \end{aligned}$$

where $(\boldsymbol{\alpha}_j)_k = -(\mathbf{x}_{gk}^T - \mathbf{x}_{gK}^T)$, $k = 1, \dots, K-1$, and we have used the fact that $\boldsymbol{\alpha}_j^T A_j^{-1} \boldsymbol{\alpha}_j = \text{Var}(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j)$. Under H_0 , $\text{Var}(\mathbf{X}_g | \mathbf{X} = \mathbf{x}_j)$ can be estimated by the sample variance of complete observations $i \in \mathcal{C}$.

B.2 The likelihood ratio test

For the EPS-full design, we use the likelihood ratio test for testing for gene-environment interactions $H_0 : \beta_{eg} = 0$. Let $\hat{\boldsymbol{\theta}}$ denote the maximum likelihood estimator of the parameters $\alpha, \beta_e, \beta_g, \beta_{eg}, \sigma$ under the alternative hypothesis, and let $\hat{\boldsymbol{\theta}}_0$ denote the corresponding maximum likelihood estimator under the null model ($\beta_{eg} = 0$). Note that under both the null and the alternative, the MLEs must be found by optimizing the EPS-full log-likelihood. The likelihood ratio test statistic is then given by

$$\lambda = 2(l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)),$$

and under the null hypothesis λ is χ^2 -distributed.

C Confounding effects in the HUNT data

We use the VO₂ data to illustrate the issue of confounding effects in the EPS-full likelihood specifically, and association studies in general. This also highlights the importance of developing models and tests for genetic effects where non-genetic environmental covariates \mathbf{x}_e are included. In the VO₂ data, we observed confounding effects between genotypes of some SNPs, sex and VO₂. Some of the SNPs in the data have genotypes that are highly correlated with sex. Sex has a strong effect on VO₂, so that spurious results were found when sex was not included as a covariate in the regression model. The effect of sex on oxygen uptake was then mediated through the genetic

variants. We have illustrated this in Figure 4. The plot Full 1 shows the $-\log_{10}(p)$ -values for the full sample when the covariate x_{sex} was not included in the model, while the plot Full 2 shows the results when x_{sex} was included. The 30 smallest p -values for the first model are marked red in both plots. We see that including sex as a covariate removed the false positives due to confounding. In Figure 4 we have also presented results for the EPS-full model when x_{sex} was not included as a covariate and X_g was not assumed dependent upon x_{sex} (EPS-full 1), the EPS-full model when x_{sex} was included as a covariate but X_g was not assumed dependent upon x_{sex} (EPS-full 2) and finally when confounding effects were accounted for by including x_{sex} as a covariate and letting the distribution of X_g be different between men and women (EPS-full 3). The top 30 findings in the first EPS-full model are marked blue in all three plots. We see that the EPS-full model is more sensitive to false positives due to confounding, compared to the full model. In EPS-full confounding must be accounted for both as a covariate in the linear model, as well as in the (unknown) distribution of the genetic variants.

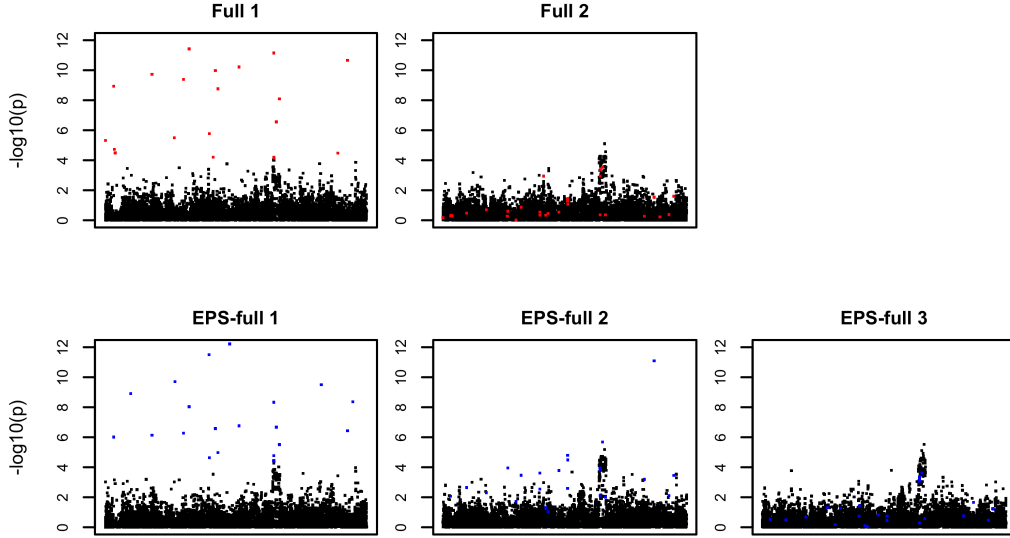


Figure 4: Manhattan-plots for testing each SNP in chromosome 1 against VO_2 for the full model (all N study participants analyzed) without including sex as a covariate in the regression model (Full 1) and when including sex as a covariate in the regression model (Full 2). Red points illustrate top 30 findings in Full 1. Manhattan-plots for the same test in the EPS-full setting (N participants analyzed of which only $n = N/2$ most extreme participants had observed genotypes) without including sex as a covariate in the regression model or accounting for sex in models of genotype distributions (EPS-full 1), including sex as a covariate in the regression model but not accounting for sex in models of genotype distributions (EPS-full 2), and including sex as a covariate in the regression model and accounting for sex in models of genotype distributions (EPS-full 3). Blue points illustrate top 30 findings in EPS-full 1.

D Power estimates under MCAR

Under the assumption that n out of N individuals in a population can be genotyped, a reasonable alternative to extreme sampling is random sampling. Throughout this paper, we compared the power of different tests for both extreme and random samples with equally many genotyped individuals. We considered random samples of size n where no information was known for the $N - n$ individuals that were not sampled for genotyping. However, as with extreme samples, random samples (RS) can also come in two types. We refer to these as RS-only and RS-full. The RS-only sample consists of observations $(y_i, \mathbf{x}_{\text{ei}}, \mathbf{x}_{\text{gi}})$ for the genotyped individuals that were randomly

β_g	RS-only	RS-complete	σ	RS-only	RS-complete	q	RS-only	RS-complete
0.3	36.18	36.26	6	76.43	76.51	0.1	42.16	42.28
0.5	76.03	76.08	8	51.82	51.92	0.2	65.61	65.73
0.7	96.34	96.37	10	36.16	36.20	0.3	76.09	76.14

Table 6: Estimated power to detect a non-null genetic effect ($H_0 : \beta_g = 0$) in simulation model (5) for different values of β_g , σ and minor allele frequency q .

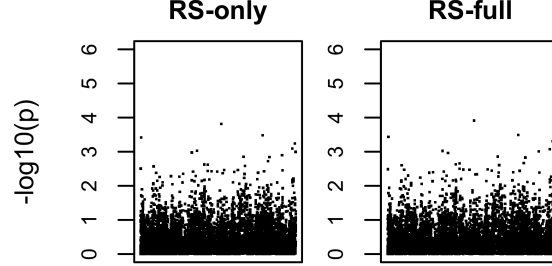


Figure 5: Manhattan-plot for testing each SNP in chromosome 1 against VO_2 in two random sampling models; RS-only ($n = N/2$ randomly drawn participants analyzed), and RS-full (N participants analyzed of which only $n = N/2$ randomly drawn participants had observed genotypes).

chosen, as used in our analysis. This sample can be analyzed with standard methods for linear regression due to the MCAR (missing completely at random) criterion. The RS-complete sample consists of observations $(y_i, \mathbf{x}_{ei}, \mathbf{x}_{gi})$ for the randomly chosen genotyped individuals, as well as observations (y_i, \mathbf{x}_{ei}) for the remaining individuals. For this sample the EPS-full likelihood can be used for model inference.

In section 5.1 of the main paper we considered extreme exposure sampling. Then the interest was on the interaction term $x_{e2}x_g$ and n individuals were genotyped due to extreme values of x_{e2} . Since x_{e2} can be regarded as a constant for the models considered here, the missing-mechanism is also here MCAR. Therefore, the following discussion can explain why the EES-full design was found to be only slightly better than the EES-only design.

White and Carlin [2010] considered a regression model with several covariates, where one covariate (X_1) had a MCAR structure. They considered the difference between a complete case (CC) approach (only individuals with observations of X_1 were analyzed) and a maximum likelihood (ML) approach that took into account all available information. They showed that for estimating the coefficient β_1 of X_1 , the variance in the estimate $\hat{\beta}_1$ was lower in the ML setting than in the CC setting, when the partial correlation of Y and X_1 given all other covariates was nonzero, i.e. when the covariate X_1 was independently associated with Y . Then the power to detect a non-zero effect of some SNP should be greater in an RS-full sample than in an RS-only sample. We evaluated this in our main-effects simulation model (5). Using the same simulation set-ups as in the main paper, the power estimates for RS-only and RS-complete were almost identical, as shown in Table 6. The RS-full method was only marginally more powerful than the RS-only method. Recall that the genetic effect size (β_g) was chosen to be very low in our simulations. For a particular data set simulated from model (5), the partial correlation of y and x_g , given x_{e1} and x_{e2} was 0.05, while the partial correlation of y and x_{e1} , given x_g and x_{e2} was 0.64. Because the partial correlation of the genetic variant and the response is close to zero, by the results of White and Carlin [2010] the difference between a method that only includes the complete cases (RS-only) and a method that includes all cases (RS-complete) is negligible. To check this results in real data, we test the VO_2 -data using the RS-only and RS-full methods. The corresponding Manhattan plots are presented in Figure 5. We see that the results from RS-full method are very similar to the RS-only method.

References

- Stian Thoresen Aspenes, Javaid Nauman, TI Nilsen, Lars Johan Vatten, and Ulrik Wisloff. Physical activity as a long-term predictor of peak oxygen uptake: the hunt study. *Med Sci Sports Exerc*, 43(9):1675–1679, 2011.
- Ian J Barnett, Seunggeun Lee, and Xihong Lin. Detecting rare variant effects using extreme phenotype sampling in sequencing association studies. *Genetic epidemiology*, 37(2):142–151, 2013.
- Thea Bjørnland, Mette Langaas, Valdemar Grill, and Ingrid Løvold Mostad. Assessing gene-environment interaction effects of FTO and MC4R with lifestyle factors on obesity using an extreme phenotype sampling design: results from the HUNT study. Submitted, 2016.
- MPM Boks, M Schipper, CD Schubart, IE Sommer, RS Kahn, and RA Ophoff. Investigating gene-environment interaction in complex diseases: increasing power by selective sampling for environmental exposure. *International journal of epidemiology*, 36(6):1363–1369, 2007.
- Stef Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45(3), 2011.
- Zehua Chen, Gang Zheng, Kaushik Ghosh, and Zhaohai Li. Linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, 77(4):661–669, 2005.
- A Darvasi and M Soller. Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. *Theoretical and applied Genetics*, 85(2-3):353–359, 1992.
- Andriy Derkach, Jerald F Lawless, and Lei Sun. Score tests for association under response-dependent sampling designs for expensive covariates. *Biometrika*, page asv038, 2015.
- Peter K. Dunn and Gordon K. Smyth. Randomized quantile residuals. *J. Comput. Graph. Statist*, 5:236–244, 1996.
- Lin T Guey, Jasmina Kravic, Olle Melander, Noël P Burt, Jason M Laramie, Valeriya Lyssenko, Anna Jonsson, Eero Lindholm, Tiinamaija Tuomi, Bo Isomaa, et al. Power in the phenotypic extremes: a simulation study of power in discovery and replication of rare variants. *Genetic epidemiology*, 35(4):236–246, 2011.
- Joel N Hirschhorn, Kirk Lohmueller, Edward Byrne, and Kurt Hirschhorn. A comprehensive review of genetic association studies. *Genetics in Medicine*, 4(2):45–61, 2002.
- BE Huang and DY Lin. Efficient association mapping of quantitative trait loci with selective genotyping. *The American Journal of Human Genetics*, 80(3):567–576, 2007.
- Joseph G Ibrahim, Ming-Hui Chen, Stuart R Lipsitz, and Amy H Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.
- Michael G Kenward and Geert Molenberghs. Likelihood based frequentist inference when data are missing at random. *Statistical Science*, pages 236–247, 1998.
- S Krokstad, A Langhammer, K Hveem, TL Holmen, K Midthjell, TR Stene, G Bratberg, J Heggland, and J Holmen. Cohort profile: the hunt study, norway. *International journal of epidemiology*, 42(4):968–977, 2013.
- Eric S Lander and David Botstein. Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199, 1989.
- JF Lawless, JD Kalbfleisch, and CJ Wild. Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(2):413–438, 1999.

- RJ Lebowitz, M Soller, and JS Beckmann. Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics*, 73(4):556–562, 1987.
- Dalin Li, Juan Pablo Lewinger, William J Gauderman, Cassandra Elizabeth Murcray, and David Conti. Using extreme phenotype sampling to identify the rare causal variants of quantitative traits in association studies. *Genetic epidemiology*, 35(8):790–799, 2011.
- Dan-Yu Lin, Donglin Zeng, and Zheng-Zheng Tang. Quantitative trait analysis in sequencing studies under trait-dependent sampling. *Proceedings of the National Academy of Sciences*, 110(30):12247–12252, 2013.
- Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2002.
- Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.
- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL <https://www.R-project.org>.
- Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.
- Pak C Sham and Shaun M Purcell. Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15(5):335–346, 2014.
- Montgomery Slatkin. Disequilibrium mapping of a quantitative-trait locus in an expanding population. *The American Journal of Human Genetics*, 64(6):1765–1773, 1999.
- Yongqiang Tang. Equivalence of three score tests for association mapping of quantitative trait loci under selective genotyping. *Genetic epidemiology*, 34(5):522–527, 2010.
- Ran Tao, Donglin Zeng, Nora Franceschini, Kari E North, Eric Boerwinkle, and Dan-Yu Lin. Analysis of sequence data under multivariate trait-dependent sampling. *Journal of the American Statistical Association*, 110(510):560–572, 2015.
- S van Buuren. Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219, 2007.
- Sofie Van Gestel, Jeanine J Houwing-Duistermaat, Rolf Adolfsson, Cornelia M van Duijn, and Christine Van Broeckhoven. Power of selective genotyping in genetic association analyses of quantitative traits. *Behavior genetics*, 30(2):141–146, 2000.
- Chris Wallace, Juliet M Chapman, and David G Clayton. Improved power offered by a score test for linkage disequilibrium mapping of quantitative-trait loci by selective genotyping. *The American Journal of Human Genetics*, 78(3):498–504, 2006.
- Mark A Weaver and Haibo Zhou. An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association*, 100(470):459–469, 2005.
- Ian R White and John B Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Statistics in medicine*, 29(28):2920–2931, 2010.
- Chao Xing and Guan Xing. Power of selective genotyping in genome-wide association studies of quantitative traits. In *BMC proceedings*, volume 3, page 1. BioMed Central, 2009.
- Haibo Zhou, Mark A Weaver, J Qin, MP Longnecker, and MC Wang. A semiparametric empirical likelihood method for data from an outcome-dependent sampling scheme with a continuous outcome. *Biometrics*, 58(2):413–421, 2002.
- Andreas Ziegler, Inke R König, and Friedrich Pahlke. *A Statistical Approach to Genetic Epidemiology: Concepts and Applications, with an E-learning platform*. John Wiley & Sons, 2010.